(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113823293 B
(45)授权公告日 2024.04.26

(21)申请号 202111140239.5
(22)申请日 2021.09.28
(65)同一申请的已公布的文献号
申请公布号 CN 113823293 A
(43)申请公布日 2021.12.21
(73)专利权人 武汉理工大学
地址 430070 湖北省武汉市洪山区珞狮路122号
(72)发明人 熊盛武 张欣冉
(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
专利代理师 罗飞
(51)Int.Cl.
G10L 17/04(2013.01)
G10L 17/18(2013.01)
G10L 17/02(2013.01)
G10L 21/0216(2013.01)
G10L 21/0232(2013.01)
G10L 25/24(2013.01)

(56)对比文件
CN 102568472 A,2012.07.11
CN 109147810 A,2019.01.04
CN 109326302 A,2019.02.12
CN 109410974 A,2019.03.01
CN 109524020 A,2019.03.26
CN 110428849 A,2019.11.08
KR 20210036692 A,2021.04.05
US 2019043529 A1,2019.02.07
CA 3179080 A1,2018.03.22
CN 109712628 A,2019.05.03
CN 111785285 A,2020.10.16
CN 104157290 A,2014.11.19
CN 104835498 A,2015.08.12
CN 107464568 A,2017.12.12
CN 110299142 A,2019.10.01
CN 112820301 A,2021.05.18 (续)

审查员 毛健

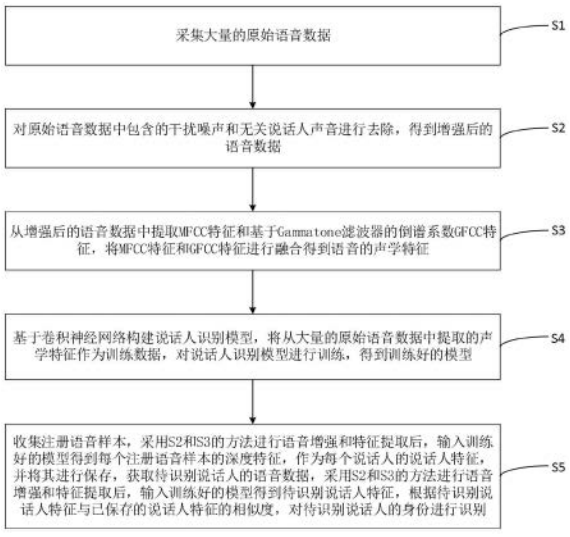
权利要求书2页 说明书8页 附图2页

(54)发明名称

一种基于语音增强的说话人识别方法及系统

(57)摘要

本发明提供了一种基于语音增强的说话人识别方法及系统,其中的方法包括如下步骤:S1采集大量的原始语音数据;S2对原始语音数据中包含的干扰噪声和无关说话人声音进行去除;S3:提取MFCC特征和GFCC特征,融合得到语音的声学特征;S4:基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练;S5:收集注册语音样本进行注册,再获取待识别说话人的语音数据,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已注册的说话人特征的相似度,对待识别说话人的身份进行识别。本发明可以提高声纹识别系统的识别准确率。



CN 113823293 B

[转续页]

[接上页]

(56)对比文件

毛维;曾庆宁;龙超.双微阵列语音增强算法在说话人识别中的应用.声学技术.2018,(第03期),全文.

王萌;王福龙.基于端点检测和高斯滤波器组的MFCC说话人识别.计算机系统应用.2016,(第10期),全文.

蔡倩等.一种基于卷积神经网络的快速说话人识别方法.《无线电工程》.2020,第50卷(第6期),第447—451页.

杨瑶;陈晓.基于神经网络的说话人识别实验设计.实验室研究与探索.2020,(第09期),全文.

蓝天;彭川;李森;叶文政;李萌;惠国强;吕忆蓝;钱宇欣;刘峤.单声道语音降噪与去混响研究综述.计算机研究与发展.2020,(第05期),全文.

1.一种基于语音增强的说话人识别方法,其特征在于,包括:

S1:采集大量的原始语音数据;

S2:对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;

S3:从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;

S4:基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;

S5:收集注册语音样本,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将其进行保存;获取待识别说话人的语音数据,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别;

其中,步骤S2采用生成对抗网络对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,实现端到端的语音增强;其中,生成对抗网络的获取方式为:将纯净语音和生活中常见噪声以随机信噪比进行混合,得到与纯净语音相对应的噪声语音,然后使用纯净语音数据集和对应的噪声语音数据集训练得到;

步骤S3包括:

S3.1:对增强后的语音数据进行语音活动端点检测,消除长时间的静音段;

S3.2:对步骤S3.1得到的语音进行预处理;

S3.3:对预处理后的语音进行快速傅里叶变换得到各帧的频谱,并对语音信号的频谱取模平方得到语音信号的功率谱;

S3.4:将快速傅里叶变换得到的功率谱通过一组梅尔尺度的三角滤波器,得到每一帧数据在三角滤波器对应频段的能量值;

S3.5:对每一帧数据在三角滤波器对应频段的能量值取对数,计算每个滤波器组输出的对数能量;

S3.6:将对数能量代入离散余弦变换,求出L阶的梅尔倒谱系数;

S3.7:将快速傅里叶变换得到的功率谱,通过Gammatone滤波器,再进行指数压缩和离散余弦变换得到语音信号的GFCC特征;

S3.8:将语音信号的MFCC特征和GFCC特征进行级联,得到语音信号的声学特征。

2.如权利要求1所述的说话人识别方法,其特征在于,步骤S1采用录音的方式进行原始语音数据的采集。

3.如权利要求1所述的说话人识别方法,其特征在于,步骤S4包括:

将大量的原始语音数据通过语音增强,然后从中提取声学特征作为训练数据,输入到说话人识别模型进行训练,得到训练好的模型。

4.如权利要求1所述的说话人识别方法,其特征在于,注册数据包括每个说话人的h个语音样本,根据待识别说话人特征与已注册的说话人特征的相似度,对待识别说话人的身份进行识别,步骤S5包括:

将注册数据中的每个语音样本进行语音增强和特征提取后,将得到的声学特征通过说

话人识别模型的卷积神经网络提取每个语音样本的深度特征;

将每个说话人的h个深度特征取平均,作为每个说话人的说话人特征,保存在数据库中;

将待识别说话人的语音数据通过语音增强和特征提取后,输入训练好的模型得到待识别说话人特征;

计算待识别说话人特征和数据库中保存的所有说话人特征的余弦相似度cos,如果最大的余弦相似度大于设定阈值,则该余弦相似度对应的数据库中的说话人即为识别到的说话人身份,否则拒绝。

5.一种基于语音增强的说话人识别系统,其特征在于,包括:

语音采集模块,用于采集大量的原始语音数据;

语音增强模块,用于对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;

语音特征提取模块,用于从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;

模型训练模块,用于基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;

说话人识别模块,收集注册语音样本,采用语音增强模块和语音特征提取模块的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将其进行保存;获取待识别说话人的语音数据,采用语音增强模块和语音特征提取模块的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别;

其中,语音增强模块采用生成对抗网络对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,实现端到端的语音增强,其中,生成对抗网络的获取方式为:将纯净语音和生活中常见噪声以随机信噪比进行混合,得到与纯净语音相对应的噪声语音,然后使用纯净语音数据集和对应的噪声语音数据集训练得到;

语音特征提取模块具体用于执行下述步骤:

S3.1:对增强后的语音数据进行语音活动端点检测,消除长时间的静音段;

S3.2:对步骤S3.1得到的语音进行预处理;

S3.3:对预处理后的语音进行快速傅里叶变换得到各帧的频谱,并对语音信号的频谱取模平方得到语音信号的功率谱;

S3.4:将快速傅里叶变换得到的功率谱通过一组梅尔尺度的三角滤波器,得到每一帧数据在三角滤波器对应频段的能量值;

S3.5:对每一帧数据在三角滤波器对应频段的能量值取对数,计算每个滤波器组输出的对数能量;

S3.6:将对数能量代入离散余弦变换,求出L阶的梅尔倒谱系数;

S3.7:将快速傅里叶变换得到的功率谱,通过Gammatone滤波器,再进行指数压缩和离散余弦变换得到语音信号的GFCC特征;

S3.8:将语音信号的MFCC特征和GFCC特征进行级联,得到语音信号的声学特征。

一种基于语音增强的说话人识别方法及系统

技术领域

[0001] 本发明涉及模式识别领域,尤其涉及一种基于语音增强的说话人识别方法及系统。

背景技术

[0002] 声纹识别,是一项提取说话人声音特征和说话内容信息,自动核验说话人身份的技术。随着人工智能在人们日常生活中的广泛应用,声纹识别技术也逐渐突显出了它的作用,比如对个人智能设备(如手机、车辆和笔记本电脑)的基于语音的认证;保证银行交易和远程支付的交易安全;以及自动身份标记。

[0003] 但是由于现实生活背景噪声的复杂,用于识别的声音总是包含着各种各样的噪声,这将会导致声纹识别效果不佳,因此如何克服待识别声音的噪声问题是声纹识别技术应用到现实生活中亟待解决的问题。

发明内容

[0004] 本发明提出一种基于语音增强的说话人识别方法及系统,用于解决或者至少部分解决现有技术中声纹识别效果不佳的技术问题。

[0005] 为了解决上述技术问题,本发明第一方面提供了一种基于语音增强的说话人识别方法,包括:

[0006] S1:采集大量的原始语音数据;

[0007] S2:对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;

[0008] S3:从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;

[0009] S4:基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;

[0010] S5:收集注册语音样本,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将其进行保存;获取待识别说话人的语音数据,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别。

[0011] 在一种实施方式中,步骤S1采用录音的方式进行原始语音数据的采集。

[0012] 在一种实施方式中,步骤S2采用生成对抗网络对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,实现端到端的语音增强。

[0013] 在一种实施方式中,步骤S3包括:

[0014] S3.1:对增强后的语音数据进行语音活动端点检测,消除长时间的静音段;

[0015] S3.2:对步骤S3.1得到的语音进行预处理;

[0016] S3.3:对预处理后的语音进行快速傅里叶变换得到各帧的频谱,并对语音信号的频谱取模平方得到语音信号的功率谱;

[0017] S3.4:将快速傅里叶变换得到的功率谱通过一组梅尔尺度的三角滤波器,得到每一帧数据在三角滤波器对应频段的能量值;

[0018] S3.5:对每一帧数据在三角滤波器对应频段的能量值取对数,计算算每个滤波器组输出的对数能量;

[0019] S3.6:将对数能量代入离散余弦变换,求出L阶的梅尔倒谱系数;

[0020] S3.7:将快速傅里叶变换得到的功率谱,通过Gammatone滤波器,再进行指数压缩和离散余弦变换得到语音信号的GFCC特征;

[0021] S3.8:将语音信号的MFCC特征和GFCC特征进行级联,得到语音信号的声学特征。

[0022] 在一种实施方式中,步骤S4包括:

[0023] 将收集的大量的原始语音数据通过语音增强,然后从中提取声学特征作为训练数据,输入到说话人识别模型进行训练,得到训练好的模型;

[0024] 在一种实施方式中,步骤S5中注册数据包括每个说话人的h个语音样本,根据待识别说话人特征与已注册的说话人特征的相似度,对待识别说话人的身份进行识别,包括:

[0025] 将注册数据中的每个语音样本进行语音增强和特征提取后,将得到的声学特征通过说话人识别模型的卷积神经网络提取每个语音样本的深度特征;

[0026] 将每个说话人的h个深度特征取平均,作为每个说话人的说话人特征,保存在数据库中;

[0027] 将待识别说话人的语音数据通过语音增强和特征提取后,输入训练好的模型得到待识别说话人特征;

[0028] 计算待识别说话人特征和数据库中保存的所有说话人特征的余弦相似度 cos,如果最大的余弦相似度大于设定阈值,则该余弦相似度对应的数据库中的说话人即为识别到的说话人身份,否则拒绝。

[0029] 基于同样的发明构思,本发明第二方面提供了一种基于语音增强的说话人识别系统,包括:

[0030] 语音采集模块,用于采集大量的原始语音数据;

[0031] 语音增强模块,用于对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;

[0032] 语音特征提取模块,用于从增强后的语音数据中提取MFCC特征和基于 Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;

[0033] 模型训练模块,用于基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;

[0034] 说话人识别模块,收集注册语音样本,采用语音增强模块和语音特征提取模块的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将每个说话人的说话人特征进行保存;获取待识别说话人的语音数据,采用语音增强模块和语音特征提取模块的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别。

[0035] 本申请实施例中的上述一个或多个技术方案,至少具有如下一种或多种技术效果:

[0036] 本发明提供的一种基于语音增强的说话人识别方法,使用端到端的语音增强方法,去除语音中的噪声和无关说话人声音,而且在声纹识别过程中使用了更加具有噪声鲁棒性的GFCC特征,并将MFCC特征和GFCC特征进行融合得到语音的声学特征,可以提高噪声鲁棒性,再基于卷积神经网络构建说话人识别模型,利用训练数据对模型进行训练,收集注册语音样本,提取每个注册说话人的说话人特征,并将其进行保存,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别。解决了现有技术中由于语音中包含的噪声而导致声纹识别效果不佳的问题,提高声纹识别的识别准确率。

附图说明

[0037] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0038] 图1为本发明实施提供的一种基于语音增强的说话人识别方法的流程图;

[0039] 图2为本发明实施中语音特征MFCC提取流程图;

[0040] 图3为本发明实施中语音特征GFCC的提取流程图;

[0041] 图4为本发明实施提供的一种基于语音增强的说话人识别系统的框图。

具体实施方式

[0042] 本发明的目的在于,提供一种基于语音增强的说话人识别方法,解决了现有技术中待识别语音中包含噪声,无法进行准确特征提取而导致的识别效果不佳的问题。

[0043] 本发明的主要构思如下:

[0044] 首先采集大量的原始语音数据,然后对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;接着从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;然后基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;收集注册语音样本,采用S2和S3 的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将其进行保存;再获取待识别说话人的语音数据,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别。

[0045] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0046] 实施例一

[0047] 本发明实施例提供了一种基于语音增强的说话人识别方法,包括:

[0048] S1:采集大量的原始语音数据;

[0049] S2:对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;

[0050] S3:从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;

[0051] S4:基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;

[0052] S5:收集注册语音样本,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将每个说话人的说话人特征进行保存;获取待识别说话人的语音数据,采用S2和S3的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的说话人特征的相似度,对待识别说话人的身份进行识别。

[0053] 具体来说,说话人识别模型训练模块中,网络模型使用卷积神经网络,分类器使用softmax,训练好的模型为离线模型。注册语音数据中包括多个说话人,每个说话人包括h个语音样本。

[0054] 请参见图1,为本发明实施提供的一种基于语音增强的说话人识别方法的流程图。

[0055] 在一种实施方式中,步骤S1采用录音的方式进行原始语音数据的采集。

[0056] 在一种实施方式中,步骤S2采用生成对抗网络对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,实现端到端的语音增强。

[0057] 生成对抗网络是一个编码器-解码器的完全卷积结构,用以去除语音中的噪声生成干净的语音波形;对抗网络在纯净语音波形和噪声语音波形基础上设定一个阈值,用于判断生成的语音波形是否纯净,当生成的语音波形和噪声语音波形的值达到该阈值时,则说明生成的语音波形已足够纯净。

[0058] 本发明在生成对抗框架内实现一种端到端语音增强方法去除语音中的干扰噪声和无关说话人声音。

[0059] 具体实施过程中,将纯净语音和生活中常见噪声以随机信噪比进行混合,得到与纯净语音相对应的噪声语音,然后使用纯净语音数据集和对应的噪声语音数据集训练得到一个实现端到端语音增强的生成对抗网络。

[0060] 下面以训练一个包含1000个纯净语音的数据集的模型为例具体说明语音模型训练过程。

[0061] 将纯净语音集和生活噪声数据集以随机信噪比(一般在-10dB至10dB之间) 进行混合,得到与纯净语音集相对应的噪声语音集。将噪声语音通过生成网络得到生成的纯净语音,然后将生成的纯净语音和真实的纯净语音通过判别网络判断生成的纯净语音是否是真实的纯净语音:如果得到的是生成的纯净语音,辨别器应该输出0,如果是真实的纯净语音应该输出1。然后通过损失函数得到误差梯度反向传播来更新参数,直到判别器无法准确判断生成的纯净语音和真实的纯净语音时,生成网络即为已经训练好的语音增强网络。直观上来说就是:判别器不得不告诉生成器如何调整从而使它生成的纯净语音变得更加真实。

[0062] 在一种实施方式中,步骤S3包括:

[0063] S3.1:对增强后的语音数据进行语音活动端点检测和消除长时间的静音段;

[0064] S3.2:对步骤S3.1得到的语音进行预处理;

[0065] S3.3:对预处理后的语音进行快速傅里叶变换得到各帧的频谱,并对语音信号的频谱取模平方得到语音信号的功率谱;

[0066] S3.4:将快速傅里叶变换得到的功率谱通过一组梅尔尺度的三角滤波器,得到每一帧数据在三角滤波器对应频段的能量值;

[0067] S3.5:对每一帧数据在三角滤波器对应频段的能量值取对数,计算算每个滤波器组输出的对数能量;

[0068] S3.6:将对数能量代入离散余弦变换,求出L阶的梅尔倒谱系数;

[0069] S3.7:将快速傅里叶变换得到的功率谱,通过Gammatone滤波器,再进行指数压缩和离散余弦变换得到语音信号的GFCC特征;

[0070] S3.8:将语音信号的MFCC特征和GFCC特征进行级联,得到语音信号的声学特征。

[0071] 具体实施过程中,预处理包括预加重、分帧和加窗。特征提取的具体步骤如下:

[0072] S301:对增强后的语音音进行语音活动端点检测(VAD),消除长时间的静音期;

[0073] S302:将语音信号通过一个高通滤波器进行预加重:$H(z)=1-\mu z^{-1}$,H(z) 为高通滤波器;μ预加重系数,通常取0.97;z为语音信号。

[0074] S303:语音信号的采样频率为16KHz,先将512个采样点集合成一帧,对应的时间长度是512/16000×1000=32ms。让两相邻帧之间有一段重叠区域,此重叠区域包含256个取样点,为采样点512的1/2。

[0075] S304:假设分帧后的信号为s(n),n=0,1,...,N-1,N为总帧数,将每一帧乘以汉明窗:

[0076] $x(n)=s(n)\times W(n)$, $W(n)=0.54-0.46\cos[\frac{2\pi n}{N-1}]$,W(n)为汉明窗;N为总帧数;n=0,1,...,N-1。

[0077] S305:对分帧加窗后的各帧信号x(n)进行快速傅里叶变换得到各帧的频谱,并对语音信号的频谱取模平方得到语音信号的功率谱。语音信号的离散傅里叶变换(语音信号以离散的形式存储)为:

[0078] $$X(k)=\sum_{n=0}^{T-1}x(n)e^{-j2\pi k/T},0\le k\le T,$$

[0079] x(n)为输入的语音信号,T表示傅里叶变换的点数。

[0080] S306:将快速傅里叶变换得到的功率谱$|X(k)|^2$通过一组梅尔尺度的三角滤波器$H_m(k)$,0≤m≤M,M为滤波器的个数:将功率谱分别跟每一个滤波器进行频率相乘累加,得到的值即为该帧数据在在该滤波器对应频段的能量值$\sum_{k=0}^{T-1}|X(k)|^2H_m(k),0\le k\le T$。

[0081] S307:对能量值取log,计算每个滤波器组输出的对数能量为:

[0082] $$s(m)=\ln(\sum_{k=0}^{T-1}|X(k)|^2H_m(k)),0\le k\le T,0\le m\le M.$$

[0083] T表示傅里叶变换的点数;M为滤波器的个数;$|X(k)|^2$为S4得到的功率谱;$H_m(k)$,0≤m≤M为一组梅尔尺度的三角滤波器。

[0084] S308:将S307的对数能量代入离散余弦变换,求出L阶的梅尔倒谱系数 MFCC:

[0085] $$C(n)=\sum_{m=0}^{T-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right),n=1,2,\ldots,L.$$

[0086] L指MFCC系数阶数,通常取12-16;M是三角滤波器个数,0≤m≤M。

[0087] S309:将快速傅里叶变换得到的功率谱,通过Gammatone滤波器,再进行指数压缩和离散余弦变换DCT得到语音信号的GFCC特征。

[0088] S310:将语音信号的MFCC特征和GFCC特征进行级联,得到语音信号的 GMCC特征。

[0089] 其中,图2和图3分别为本发明实施中语音特征MFCC提取流程图和语音特征GFCC的提取流程图。

[0090] 在一种实施方式中,步骤S4包括:

[0091] 将收集的大量的原始语音数据通过语音增强,然后从中提取声学特征作为训练数据,输入到说话人识别模型进行训练,得到训练好的模型。

[0092] 具体来说,训练模型为离线过程,说话人识别模型的训练:

[0093] 采用录音的方式收集训练样本;将收集到的语音样本通过语音预处理模块(语音增强模块和语音特征提取模块)得到语音的GMCC特征;将GMCC特征作为模型的输入,采用卷积神经网络结构和softmax分类训练说话人识别模型。

[0094] 下面以训练一个包含1000个说话人的模型为例具体说明说话人识别模型训练过程。

[0095] 采集每个说话人的样本,每人采集100个样本;将所有的语音样本通过语音预处理模块(语音增强模块和语音特征提取模块)得到语音的GMCC特征作为卷积神经网络(说话人识别模型)的训练数据,其中,将所有训练数据随机分为 5:1,分别作为训练集和验证集;使用训练集训练卷积网络,当训练过的卷积网络在验证集上的识别精度基本保持不变时,卷积网络训练完成;否则继续训练。该训练完成的卷积网络即为说话人识别离线模型。

[0096] 在一种实施方式中,步骤S5中注册数据包括每个说话人的h个语音样本,根据待识别说话人特征与已注册的说话人特征的相似度,对待识别说话人的身份进行识别,包括:

[0097] 将注册数据中的每个语音样本进行语音增强和特征提取后,将得到的声学特征通过说话人识别模型的卷积神经网络提取每个语音样本的深度特征;

[0098] 将每个说话人的h个深度特征取平均,作为每个说话人的说话人特征,保存在数据库中;

[0099] 将待识别说话人的语音数据通过语音增强和特征提取后,输入训练好的模型得到待识别说话人特征;

[0100] 计算待识别说话人特征和数据库中保存的所有说话人特征的余弦相似度 cos,如果最大的余弦相似度大于设定阈值,则该余弦相似度对应的数据库中的说话人即为识别到的说话人身份,否则拒绝。

[0101] 注册模式:

[0102] 采用录音的方式收集注册样本;将收集到的注册样本通过语音预处理模块得到语音的GMCC特征;将语音的GMCC特征通过说话人识别离线模型提取每个语音样本的Deep Feature(深度特征);生成注册数据(即每个说话人的说话人特征),存放在数据库中。

[0103] 例如,采集10个说话人的样本(每人20个语音样本);语音预处理模块处理所有语

音样本,得到语音的GMCC特征;将语音的GMCC特征通过说话人识别离线模型得到200个语音样本的Deep Feature;然后将每个说话人的20个Deep Feature取平均,作为每个说话人特征;将10个说话人特征保存在数据库中: speaker0,speaker1,.....,speaker9。

[0104] 识别模式:

[0105] 采用录音的方式收集待识别样本;将待识别样本通过语音预处理模块得到 GMCC特征;将GMCC特征通过说话人识别离线模型得到待识别样本的Deep Feature,作为待识别说话人特征;计算待识别说话人特征和数据库中的所有说话人特征的余弦相似度cos,如果最大的余弦相似度大于某个阈值,则该余弦相似度对应的数据库中的说话人即为识别到的说话人;否则拒绝。

[0106] 举例来说,采集此说话人的语音数据一条;通过语音预处理模块得到GMCC 特征;将GMCC特征通过说话人识别离线模型得到该语音数据的Deep Feature,作为此说话人特征;将此说话人特征和数据库中保存的10个说话人特征计算余弦相似度得到cos0,cos1,...,cos9,找到这10个余弦相似度中的最大值cos_max和对应说话人的编号speaker_x,如果这个最大值大于设定阈值,则接受此说话人为speaker_x,否则识别为未注册说话人。

[0107] 综上所述,本发明通过语音采集、语音增强、语音特征提取、说话人模型训练、说话人注册、说话人识别实现了一种基于语音增强的说话人识别方法。

[0108] 相对于现有技术,本发明的有益效果是:

[0109] 本发明提出的一种基于语音增强的说话人识别方法及系统,使用端到端的语音增强方法,去除语音中的噪声和无关说话人声音,而且在声纹识别过程中使用了更加具有噪声鲁棒性的GFCC特征,提高了整个系统的噪声鲁棒性,可以解决由于语音中包含的噪声而导致声纹识别效果不佳的问题,提高声纹识别系统的识别准确率。

[0110] 实施例二

[0111] 基于同样的发明构思,本实施例提供了一种基于语音增强的说话人识别系统,请参见图4,该系统包括:

[0112] 语音采集模块201,用于采集大量的原始语音数据;

[0113] 语音增强模块202,用于对原始语音数据中包含的干扰噪声和无关说话人声音进行去除,得到增强后的语音数据;

[0114] 语音特征提取模块203,用于从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征,将MFCC特征和GFCC特征进行融合得到语音的声学特征;

[0115] 模型训练模块204,用于基于卷积神经网络构建说话人识别模型,将从大量的原始语音数据中提取的声学特征作为训练数据,对说话人识别模型进行训练,得到训练好的模型;

[0116] 说话人识别模块205,用于注册说话人和识别说话人,收集注册语音样本,采用语音增强模块和语音特征提取模块的方法进行语音增强和特征提取后,输入训练好的模型得到每个注册语音样本的深度特征,作为每个说话人的说话人特征,并将其进行保存;获取待识别说话人的语音数据,采用语音增强模块和语音特征提取模块的方法进行语音增强和特征提取后,输入训练好的模型得到待识别说话人特征,根据待识别说话人特征与已保存的

说话人特征的相似度,对待识别说话人的身份进行识别。

[0117] 由于本发明实施例二所介绍的系统,为实施本发明实施例一种基于语音增强的说话人识别方法所采用的系统,故而基于本发明实施例一所介绍的方法,本领域所属技术人员能够了解该系统的具体结构及变形,故而在此不再赘述。凡是本发明实施例一的方法所采用的系统都属于本发明所欲保护的范围。

[0118] 以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

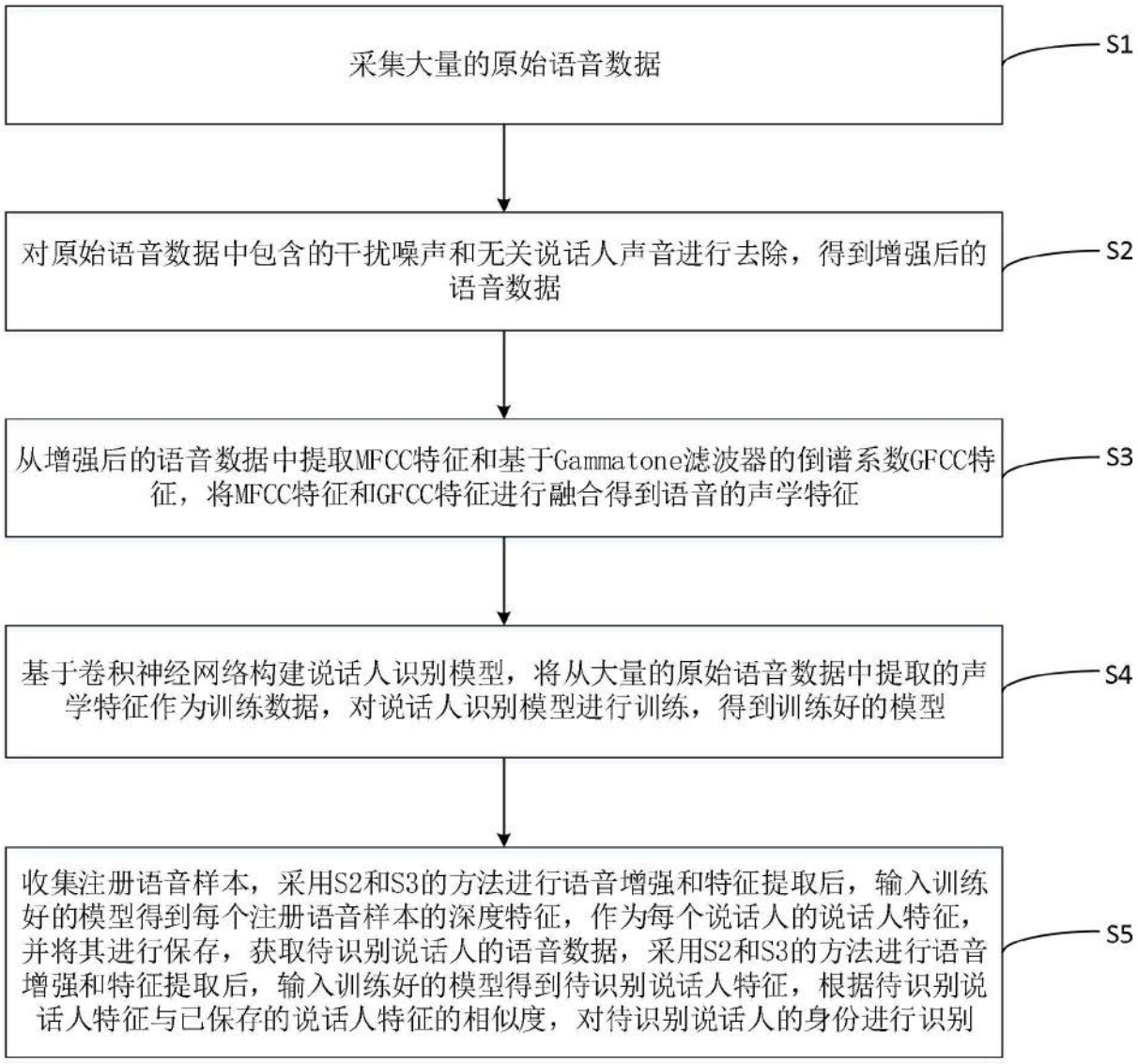

采集大量的原始语音数据
S1
对原始语音数据中包含的干扰噪声和无关说话人声音进行去除，得到增强后的语音数据
S2
从增强后的语音数据中提取MFCC特征和基于Gammatone滤波器的倒谱系数GFCC特征，将MFCC特征和GFCC特征进行融合得到语音的声学特征
S3
基于卷积神经网络构建说话人识别模型，将从大量的原始语音数据中提取的声学特征作为训练数据，对说话人识别模型进行训练，得到训练好的模型
S4
收集注册语音样本，采用S2和S3的方法进行语音增强和特征提取后，输入训练好的模型得到每个注册语音样本的深度特征，作为每个说话人的说话人特征，并将其进行保存，获取待识别说话人的语音数据，采用S2和S3的方法进行语音增强和特征提取后，输入训练好的模型得到待识别说话人特征，根据待识别说话人特征与已保存的说话人特征的相似度，对待识别说话人的身份进行识别
S5

图1

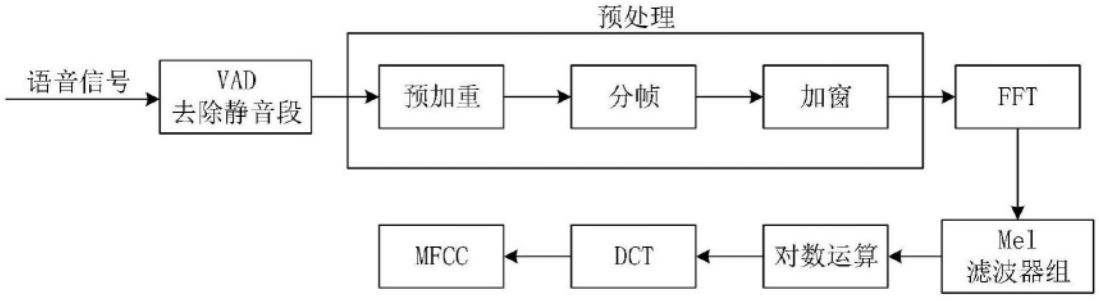

预处理
语音信号
VAD
去除静音段
预加重
分帧
加窗
FFT
Mel
滤波器组
对数运算
DCT
MFCC

图2

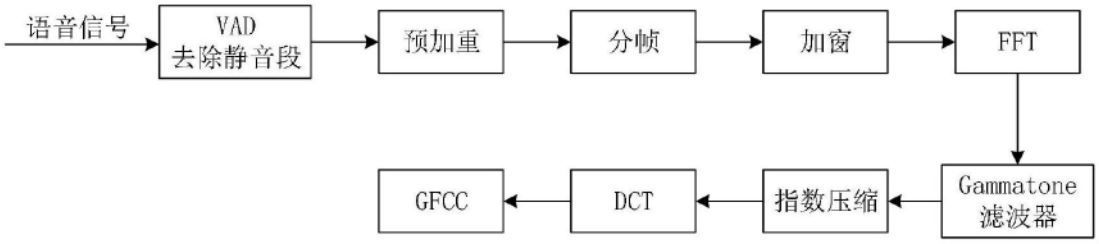

语音信号
VAD
去除静音段
预加重
分帧
加窗
FFT
Gammatone
滤波器
指数压缩
DCT
GFCC

图3

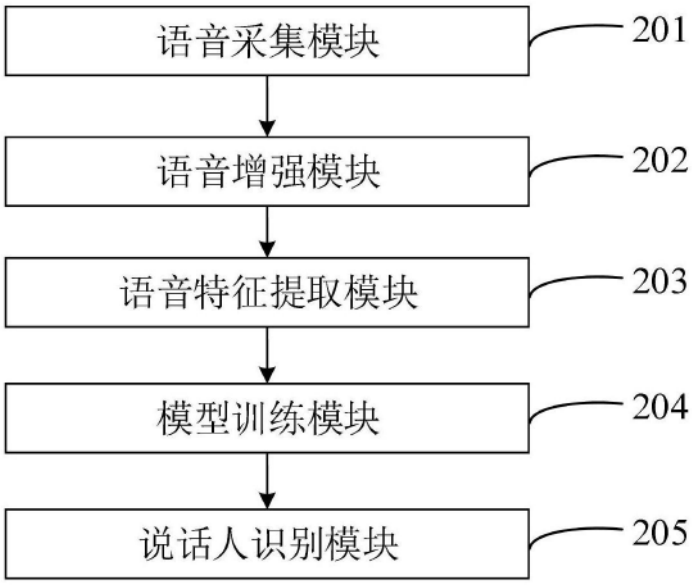

语音采集模块
201
语音增强模块
202
语音特征提取模块
203
模型训练模块
204
说话人识别模块
205

图4