基于 GMM 统计特性的电子伪装语音鉴定研究

李燕萍,林 乐,陶定元
(南京邮电大学 通信与信息工程学院,江苏 南京 210000)

摘 要:数字多媒体技术的发展使多媒体信息得到广泛使用和传播,给人类的信息交流带来极大的便利。随着语音相关技术的发展与逐渐成熟,对于语音信号处理的应用也越来越广泛。数字多媒体信息易于修改的特点,使其面临着恶意篡改带来的严重危机。近年来,手机应用软件市场上出现了大量的变声软件,例如微信变声器、超级变声器等等,类似变声器的下载量动辄上百万,这些应用软件可使说话人的声音发生巨大的改变,致使一般的听话人无法辨认发音人的身份、年龄乃至性别,即使是对话者非常熟悉的人也很难识别出说话者的身份。提出了一种鉴定电子伪装语音的方法,通过 GMM 模型建模,将其均值矢量构成组合特征,然后基于 SVM 分类器进行训练和鉴别。通过对比语音伪装前后的梅尔倒谱特征参数的统计特性变化,对特征参数的变化规律进行了分析研究。实验结果表明,提出的方法对电子伪装语音的鉴定正确率达到 90%。

关键词:变声软件;电子伪装语音;梅尔倒谱系数;支持向量机;高斯混合模型
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2017)01-0103-04
doi:10.3969/j.issn.1673-629X.2017.01.023

Research on Identification of Electronic Disguised Voice Based on GMM Statistical Parameters

LI Yan-ping, LIN Le, TAO Ding-yuan
(College of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210000, China)

Abstract: With the development of digital multimedia technology, digital information has been widely used and spread, which brings great convenience to human communication. Speech related technology gradually becomes mature, and its application is more and more extensive. This kind of information is easy to be modified, so that it is facing a serious crisis of malicious tampering. In recent years, a large number of software appear in mobile phone application store, such as Wechat Voice Changer, Super Voice Changer and so on, which can change the speaker's voice a lot. As a result, the listener cannot identify the speaker's age and sex, even they are familiar. A novel algorithm for identification of electronic disguised voice is put forward based on supervector combined by mean vectors of Gaussian mixture model and SVM classifier for training and identification. By comparing the statistical change of MFCC between nature and disguised voice, the variation of voice parameters is studied. Experimental results show that the identification rate can reach 90%.

Key words: voice changer; electronic disguised voice; MFCC; SVM; GMM

0 引 言

近年来,手机应用软件市场上出现了大量的变声软件,例如微信变声器、超级变声器等等,类似变声器的下载量动辄上百万,这些应用软件可使说话人的声音发生巨大的改变,致使一般的听话人无法辨认发音人的身份、年龄乃至性别,即使是对话者非常熟悉的人也很难识别出说话者的身份。犯罪分子利用电子伪装语音[1-3]实施电话诈骗严重危害社会安全,由于伪装语音具有良好的伪装特性,给司法鉴定工作带来很大的困难。鉴于电子伪装语音的严重危害,亟待寻求一种鉴定电子伪装语音的方法。

目前,对于电子伪装语音相关的研究,大多集中在电子伪装语音对于说话人识别系统的识别率的影响。文献[4-6]表明,电子伪装语音严重影响说话人识别

收稿日期:2015-10-26 **修回日期**:2016-02-25 **网络出版时间**:2017-01-04
基金项目:国家自然科学基金资助项目(61401227);江苏省博士后基金(1402067B);智能语音技术公安部重点实验室 2014 年度开放课题(2014ISTKFKT02)
作者简介:李燕萍(1983-),女,副教授,博士,研究方向为说话人识别、语音转换;林 乐(1990-),男,硕士研究生,研究方向为说话人识别。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20170104.1023.022.html

系统的识别率。文中提出了一种用于鉴定电子伪装语音的方法,在进行说话人识别实验之前采用该方法进行电子伪装语音鉴定,可有效提高说话人识别系统的识别率。该方法采用梅尔倒谱系数(MFCC)结合高斯混合模型(GMM)[7-9],以 GMM 模型均值组合特征向量作为 SVM 分类器训练和鉴别的特征参数[10]。经实验结果证明,这种方法对于电子伪装语音的鉴定率达到 90%。

1 电子伪装语音基本原理

电子伪装语音的基本原理是通过调整采样率即时域的压缩或者展宽从而改变基音频率[11],用这种方法可以非常简单地对改变声音,但是这样的伪装语音往往听起来不自然,有的变声软件采用基音同步叠加相加法对语音进行进一步处理,从而使伪装之后的语音听起来更加自然。

在语音学研究中,基音频率被认为是最多降低或升高 12 个半音。假设语音基音频率为 f_0,伪装因子 a 为提高或降低 a 个半音,经过伪装之后的基音频率为 f_1,则有:

$$f_1 = 2^{(a-1)/12} \cdot f_0 \tag{1}$$

其中,伪装因子 a 为整数,且 $-11 < a < 13$。如果 $a > 1$,说明提高了基音频率;如果 $a < 1$,说明降低了基音频率;如果 $a = 1$,说明未改变基音频率。

2 鉴别方法原理介绍

2.1 特征参数提取过程

语音的预处理包括端点检测、预加重、分帧、加窗。假设一段语音 $x(n)$,经过预处理之后,得到 N 帧语音,对这 N 帧语音提取 D 维 MFCC 系数,得到 N 个 D 维向量,用这 N 帧训练高斯混合模型。

为了准确地表征说话人的个性特征,往往需要说话人大量的样本,而将大量的样本输入到支持向量机进行分类时,会有巨大的计算量,自然而然地,通过少量的样本作为支持向量机的输入,选取代表性样本的方法有很多。例如,对于 MFCC 或 LPCC 特征向量序列可以通过随机方式、矢量量化等方法选取,但是这些方法具有很明显的缺点。随机选取的样本由于具有很强的偶然性,难以表示大量样本的分布情况,而矢量量化方法虽能很好地表示样本的分布中心,但仍包含很多冗余信息,并且鲁棒性较差[12]。

GMM 模型作为一种统计模型,利用若干高斯概率密度函数的加权和来表示特征向量在概率空间的分布情况,GMM 模型使用较少的参数很好地描述了说话人的个性特征,在文本无关说话人识别方面得到了广泛应用[13]。GMM 模型由 EM 算法训练得到,其均值向量不但反映了各说话人在特征空间的分布,而且也较好地反映了说话人的个性信息,因而可考虑采用 GMM 模型的均值向量作为 SVM 的训练样本。一个具有 M 个混合数 D 维 GMM 可表示为:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{2}$$

其中,$\vec{x}$ 为观测向量;p_i 为每个高斯分量的加权值,且满足条件 $\sum_{i=1}^{M} p_i = 1$;$b_i(\vec{x})$ 为每个高斯分量的概率密度函数,表示如下:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{ -\frac{1}{2} (\vec{x} - V_i)' \boldsymbol{\Sigma}_i^{-1} (\vec{x} - V_i) \right\} \tag{3}$$

其中,$\boldsymbol{V}_i$ 为均值矢量;$\boldsymbol{\Sigma}_i$ 为协方差矩阵;λ 为模型参数,$\lambda = \{p_i, \boldsymbol{V}_i, \boldsymbol{\Sigma}_i\}$,$i = 1,2,\cdots,M$。

则有均值组合特征向量:

$$\boldsymbol{V} = (V_1, V_2, \cdots, V_M) \tag{4}$$

其中,V_i,$i = 1,2,\cdots,M$ 为第 i 个高斯混合模型均值向量。

2.2 分类算法

文中采用的分类算法是建立在 MFCC 统计特性的基础上,训练和鉴别流程分别见图 1 和图 2。

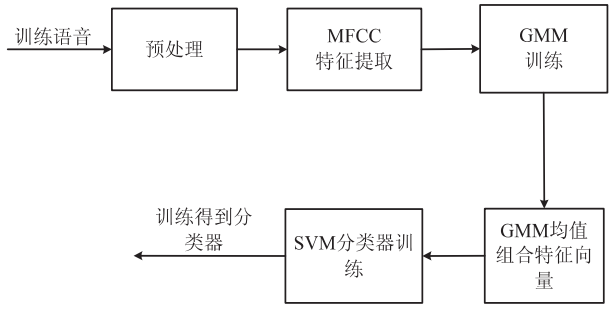


图 1 训练流程图

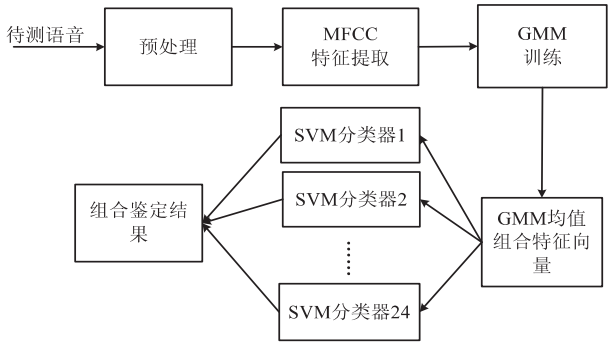


图 2 鉴别流程图

在训练阶段,训练数据库由原始语音和伪装语音数据集组成,根据伪装因子 a 可分为 24 个子集,将每个子集中的伪装语音和原始语音进行预处理及 MFCC

特征提取,然后训练 GMM 模型,得到均值组合特征向量 $\boldsymbol{V}$,以该特征参数作为输入样本进行 SVM 训练,从而得到其中一个子集的分类器。同理,可以训练得到 24 个分类器[14]。

在识别阶段,待测试语音进行同样的预处理和 MFCC 特征提取,训练 GMM 模型,得到均值组合特征向量 $\boldsymbol{V}$,将该特征向量分别输入到 24 个分类器中,如果 24 个分类器全部鉴别为原始语音,则判定待测试语音为原始语音,只要其中一个鉴别为伪装语音,则判定为伪装语音。

3　实验结果与分析

实验用的变声软件是一款非常典型的手机变声软件,软件名字叫“高保真变声”,实验用的手机系统平台是 Android4.2,录音的采样频率为 8 kHz,PCM 方式量化精度为 16 bit,语料人数是 24 人,其中训练语音 240 段,测试语音 240 段,语音内容包括“你好”、“快把钱给我”、“把钱转到我的银行卡里”、“你的小孩在我手里,你赶快拿钱来赎”、“南京邮电大学”等 1 ~ 10 s 长短不一的语音。

3.1　语音的预处理

首先对读取的语音输入信号进行端点检测,去除静音段,对语音进行预加重,预加重的目的在于滤除低频干扰,将更为有用的高频部分的频谱进行适当提升,文中实验采用的预加重系数为 0.98。然后进行分帧加窗处理,实验提取 MFCC 参数时选取的帧长为 20 ms,帧移为 10 ms,分帧之后加汉明窗进行处理。

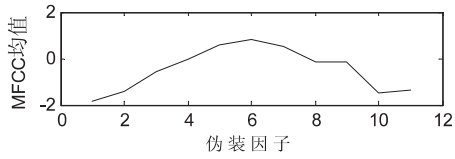
3.2　语音的特征参数提取

对经过预处理的语音提取 20 维 MFCC,图 3 列出了部分男声“Hello”的 MFCC 统计特性变化,其中伪装程度为 1(即 $a = 1$)时为原始语音的统计特性。

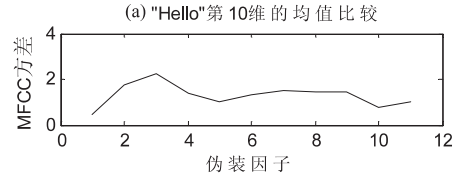
图中对比了第 10、17 维 MFCC 系数统计特性的变化,从中可以看出,第 10 维 MFCC 系数均值先递增而后递减,方差先递增而后平稳波动。第 17 维 MFCC 系数随着伪装因子的提高先增加而后降低再增加,方差在伪装因子为 5 和 7 时取得极大值。

从上面的分析比较可以得出以下结论:无论伪装因子为何值,伪装语音与原始语音的 MFCC 系数的统计均值和方差均有较大差异。这种统计特性的差异为选取 GMM 均值组合特征参数奠定了理论基础。

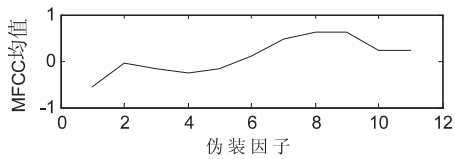
在选取高斯混合分量个数时,个数越多,对说话人特征矢量空间分布就越逼近,从而提高了系统的鲁棒性。但是高斯混合分量个数太多,一方面加大了系统的计算、占用更多的系统资源,另一方面,在有限时长的训练数据情况下使得模型训练不够充分,从而降低了系统性能。一般认为模型的高斯混合分量个数在 32 以上系统的性能趋于稳定。所以在提取得到语音的 MFCC 系数后,选取高斯混合模型个数为 48,进行 GMM 训练,从而得到均值组合特征向量 $\boldsymbol{V}$。文中对比了不同伪装因子 V_2、V_4 两个高斯模型分量均值,如图 4 所示。



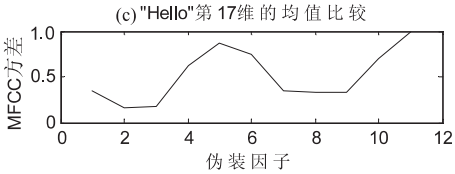
(a) "Hello"第 10维的均值比较



(b) "Hello"第 10维的方差比较

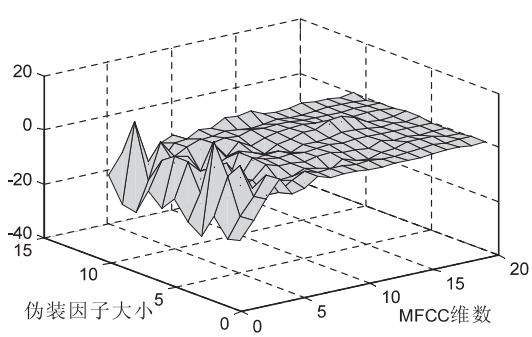


(c) "Hello"第 17维的均值比较

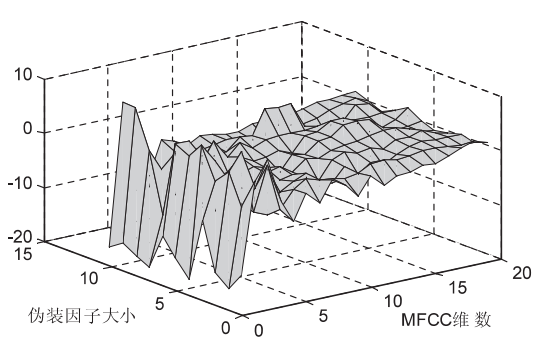


(d) "Hello"第 17维的方差比较

图 3　MFCC 统计特性变化比较



(a) 不同伪装因子高斯混合均值 V_2 比较



(b) 不同伪装因子高斯混合均值 V_4 比较

图 4　GMM 模型均值向量比较

由图 4 可知,不同的伪装因子下 V_2、V_4 两个高斯混合模型均值同样具有较大差异。

3.3 分类器的训练及伪装语音的鉴定

文中利用语音高斯混合模型参数混合均值构成的组合向量作为 SVM 分类器的输入进行训练,训练得到伪装因子从-11 ~ 13 的 24 个分类器,待测语音从 2 ~ 60 s 不等,待测伪装语音伪装因子分布从-12 ~ 12,每个伪装因子取 10 段语音待测。实验结果见表 1。

表 1 不同伪装因子的鉴别率结果

伪装因子 a	鉴别率/%
-11	80
-10	100
-9	80
-8	100
-7	100
-6	80
-5	100
-4	80
-3	100
-2	90
-1	90
0	80
2	80
3	70
4	90
5	100
6	80
7	100
8	90
9	100
10	90
11	100
12	80
13	90

从表中可以看出,在伪装程度很小时,伪装语音鉴定的正确率比伪装程度较大时要低,这是因为伪装语音与原始语音差别并不明显,但尽管这种差异不明显,鉴定正确率也可达到 80% 。综合上述数据,对伪装语音的总体鉴定正确率为 90% 。

4 结束语

电子伪装语音对于说话人识别识别系统的识别率有较大影响。为了去除电子伪装语音的不良影响,提出了一种 SVM 结合 GMM 均值组合特征参数的电子伪装语音鉴定方法。并且运用这种方法有效地实现了对电子伪装语音的鉴定,在进行说话人识别实验之前,采用该方法对语音进行电子伪装语音鉴定,有效提高了说话人识别系统识别率。实验结果表明,该方法鉴别电子伪装语音的效果可达 90% 。因此该方法可为将来电子伪装语音的司法鉴定提供理论依据。鉴于实验人数较少,测试语音说话人来自训练语音说话人集合,在今后的工作中,会使用不同的语料库,从而实现对不是来自训练语音说话人集合测试语音的鉴定。

参考文献:

[1] Neustein A, Patil H A. Forensic speaker recognition: law enforcement and counter - terrorism [M]. [s. l.]: Springer, 2011.

[2] Wu Haojun, Wang Yong, Huang Jiwu. Blind detection of electronic disguised voice [C]//Proceedings of the international conference on acoustic, speech and signal processing. [s. l.]: [s. n.], 2013: 3013-3017.

[3] 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.

[4] Zhang C, Tan T J. Voice disguise and automatic speaker recognition [J]. Forensic Science International, 2008, 175 (2-3): 118-122.

[5] Hermann J K, Joaquin G R, Javier O G. Effect of voice disguise on the performance of a forensic automatic speaker recognition system [C]//Proceedings of the speaker and language recognition workshop. [s. l.]: [s. n.], 2014.

[6] Rodman R. Computer recognition of speakers who disguise their voice [C]//Proceedings of the international conference on signal processing applications & technology. Texas: [s. n.], 2000: 474-476.

[7] Kinnunen T, Li Haizhou. An overview of text-independent speaker recognition: from features to supervectors [J]. Speech Communication, 2010, 52(1): 12-40.

[8] 于 明,袁玉倩,董 浩,等. 一种基于 MFCC 和 LPCC 的文本相关说话人识别方法[J]. 计算机应用,2006,26(4):883-885.

[9] 蒋 晔,唐振民. GMM 文本无关的说话人识别系统研究[J]. 计算机工程与应用,2010,46(11):179-182.

[10] 冷自强,王金明,林大会. 一种 GMM-SVM 混合说话人辨认模型[J]. 军事通信技术,2009,30(1):86-89.

[11] Trehub S E, Cohen A J, Thorpe L A, et al. Development of the perception of musical relations: semitone and diatonic structure [J]. Journal of Experimental Psychology-human Perception and Performance, 1986, 12(3): 295-301.

[12] 林 琳,陈 虹,陈 建,等. 基于多核 SVM-GMM 的短语音说话人识别[J]. 吉林大学学报:工学版,2012,43(2):504-509.

[13] 贺志阳,张玲华. 基于 GMM 统计参数和 SVM 的说话人辨认研究[J]. 南京邮电大学学报:自然科学版,2006,26(3):78-82.

[14] Wu J H, Wang Yong. Identification of electronic disguised voices [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(3): 489-499.