文章编号: 1004-4353(2017)02-0184-05

改进的跨语种说话人确认方法的研究

朱虹, 金小峰*
(延边大学工学院 计算机科学与技术学科 智能信息处理研究室, 吉林 延吉 133002)

摘要: 提出了一种基于改进的语音融合特征和 GMM 模型相结合的跨语种说话人确认方法. 首先,采用 Teager 能量算子提取语音中的浊音段,消除与说话人声道特征无关的静音段和清音段. 其次,提取基音周期参数,并与 16 维的 MFCC 参数融合形成本文的语音融合特征. 最后,将本文方法与文献[9]的方法分别进行了单语种和跨语种的说话人确认对比实验,实验结果表明本文方法识别准确率和平均判别时间均优于文献[9]的方法,证明本文提出的方法有效,可用于跨语种的说话人确认应用领域.

关键词: 说话人确认; 跨语种; 浊音段提取; 融合特征
中图分类号: TP391.41 **文献标识码**: A

Research of improved cross-lingual speaker verification method

ZHU Hong, JIN Xiaofeng*
(*Intelligent Information Processing Lab., Dept. of Computer Science & Technology, College of Engineering, Yanbian University, Yanji* 133002, *China*)

Abstract: This paper presents a cross-lingual speaker verification method based on improved speech fusion feature and GMM model. First, the Teager energy operator is used to extract voiced clips in speech, eliminating mute and unvoiced clips that are independent of speaker's vocal tract. Secondly, pitch period parameters are extracted and fused with 16-dimensional MFCC parameters to form speech fusion feature. Finally, experimental results show that the accuracy and average discriminant time of this method are better than that of reference [9], which proves that the method proposed in this paper is valid and available in cross-lingual speaker verification applications.

Keywords: speaker verification; cross-lingual; voiced extraction; fusion feature

0 引言

在国际化潮流以及多民族和多文化相互交融的时代背景下,人们使用和交流的语言不止一种,基于多语种的说话人识别成为必须要解决的问题[1]. 说话人识别技术大致分为 3 类[2]: ① 说话人确认和说话人辨认; ② 与文本有关的和无关的说话人识别; ③ 单语种、跨语种、多语种的说话人识别. 其中单语种指的是测试阶段和训练阶段使用相同语种的语音;跨语种指训练语音是某一语种,测试语音是另外一个语种;多语种指训练语音为某一语种,测试语音包含混合的多语种. 本文研究的是与文本无关的、跨语种的说话人确认方法.

跨语种和多语种的说话人识别和确认方法在国内外已经取得了一定的成果,例如:Sarkar 等[3]使用 IITKGP-MLILSC 语料库,采用 MFCC 特征参数和 GMM 的算法针对 13 种印度语言进行了闭集的

收稿日期: 2017-04-07 **基金项目**: 吉林省科技厅自然科学基金资助项目(20140101225JC)
* **通信作者**: 金小峰(1970—),男,教授,研究方向为智能信息处理.

多语种说话人确认实验，平均同等错误率(EER)为 11.71%；Bhattacharjee 等[4]针对英语、印地语和一种阿鲁纳恰尔邦的当地语言等 3 种语言，使用 MFCC 特征参数和基于 GMM-UBM 模型算法，对训练阶段和测试阶段中的语言失配因素进行了评估；Ma 等[5]针对英语-汉语，在建立说话人模型时采用双语的语音，训练出双语种的说话人模型，得到了较好的识别效果. 相比之下，国内多语种说话人识别技术的发展仍然滞后.

跨语种说话人识别的主要难点在于每个语种都携带着其语言信息中特殊的语言因素，比如音素和声调以及发音时发音器官的张驰程度等等，这些因素的差异在一定程度上会对实验结果造成影响[6]. 因此，本文在基于人耳听觉特征的 MFCC 特征参数上添加一个与人类声道信息相关的基音周期参数. 基音周期与说话人声带的长短、韧性和发音习惯有关，反映了说话人生理上的特征差异. 虽然基音周期的变化与文本有关，然而从统计意义上来说，不同说话人基音周期分布具有一定的差异性；因此，基音周期可用于文本无关的说话人识别[7].

MFCC 参数相对稳定但易被模仿，且提取过程没有考虑发音的过程[8]. 清音和静音段与浊音段不同，不能提供足够的说话人个性信息，因此本文提取语音特征时先剔除了静音和清音，仅对浊音段的基音周期和加权 MFCC 参数进行融合. 与文献[9]的方法相比，本文方法保留了加权 MFCC 的前 16 维(文献[9]为 39 维)，因此减少了计算量，同时因为剔除了静音和清音提高了说话人确认的准确率.

1　融合特征的提取方法

1.1　浊音段的提取与整合

在语音信号时域短时段(10～30 ms)时间内信号近似稳定，区分清音和浊音常利用短时能量、短时幅度、短时过零率、线性预测编码参数等方法来进行，但仍然存在时变的成分[10]，而 Teager 能量算子能够有效地追踪信号的瞬间能量，可用于语音中的时变信号分析[11]. Teager 能量算子(Teager energy operator，TEO)是由美国科学家 Teager 在研究非线性语音建模时提出的一种非线性算子，对于有限频带的信号 $x(n)$此算子可以近似表示为

$$\psi[x(n)]=[x(n)]^2-x(n+1)x(n-1). \tag{1}$$

Teager 能量算子提取包络线是对被测波形相邻的 3 个采样点进行计算，具有优良的时间分辨率，且简单、快速，对于语音信号中的时变部分能实时跟踪其波形变化[12-13]. 提取和整合浊音段的步骤如下：

Step 1　预处理. 滤除 50 Hz 直流噪声和大于 4 kHz 的高频噪声，语音信号 $s(n)$ 通过 60～4 kHz 频率的带通滤波器 $H(n)$ 得到：

$$x(n)=s(n)\times H(n),\ n=1,2,\cdots,N, \tag{2}$$

其中 N 是语音信号的帧长.

Step 2　快速傅里叶变换. 对 $x(n)$ 进行快速傅里叶变换，

$$X(k)=\sum_{n}^{N-1}x(n)\mathrm{e}^{-\frac{\mathrm{j}2\pi(k-1)(n-1)}{N}},\ 1\leqslant k\leqslant N. \tag{3}$$

Step 3　计算 TEO 非线性算子. 采用公式(1) 可得到如下 TEO 算子：

$$t(k)=\psi[X(k)]. \tag{4}$$

Step 4　选取浊音段 $v(k)$. 由于浊音 $t(k)$ 大于清音和静音，因此可由阈值 h 来提取浊音段：

$$v(k)=\begin{cases}1, & t(k)\geqslant h;\\ 0, & t(k)<h.\end{cases} \tag{5}$$

本文取阈值 $h=0.02$.

Step 5　获取浊音段对应的语音帧号，将浊音段串联在一起，整合为新的语音段，作为提取融合特征的原始语音.

1.2 融合特征

首先通过计算语音信号的 TEO 值去除静音段和清音段,将获取的浊音段整合为新的语音段;然后提取新语音段中的基音周期和加权的 WMFCC 参数[14];最后将得到的 1 维基音周期和 16 维的 WMFCC 串联成新的特征参数形成融合特征.融合特征提取过程如图 1 所示.

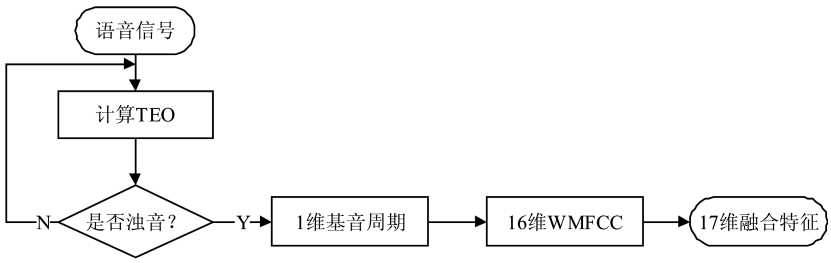

图 1 融合特征提取流程图

人的声道(含口腔、鼻腔)特征因人而异,作为区分声道特性的语音特征基音周期参数能够很好地区分不同的说话人.另外,MFCC 是模仿人耳听觉特性的特征参数,本文采用的听觉特性联合声道特性能够更全面地描述每个说话人的语音特征.

2 跨语种说话人确认算法

高斯混合模型(GMM)常被作为跨语种说话人识别的主要特征模型,通常利用 K-means 聚类方法和 EM 算法训练得到说话人的 GMM 模型.本文在测试阶段采用最大似然函数法计算测试语音与 GMM 模型的得分,然后再与阈值进行对比得出判决结果.图 2 为本文实现跨语种说话人识别的具体流程(训练语音是英语、朝鲜语、蒙古语、日语的其中一个,测试语音采用汉语).

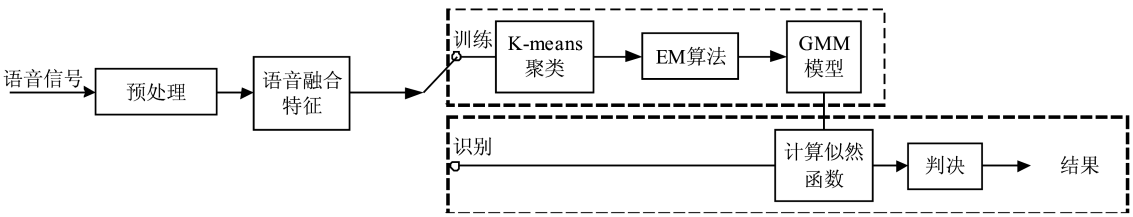

图 2 跨语种说话人确认系统框图

根据图 2,本文提出的跨语种说话人确认学习算法如下:

输入:N 个测试者英语(或朝鲜语、蒙古语、日语其中之一)语音段 S, $S=\{s_i\}$, $i=1,2,\cdots,N$; $s_i=\{s_i^{(l)_j}\}$, $j\geqslant 0$, $l\in\{\text{English},\text{Korean},\text{Japenese},\text{Mongolian}\}$.

Step 1 利用本文提出的融合特征的提取方法计算和提取每个说话人多语种语音的 17 维融合特征参数, $F=\{F_1,F_2,\cdots,F_i,\cdots,F_N\}$; $F_i=\{f_1,\cdots,f_T\}$,其中 T 为帧数.

Step 2 通过 K-means 聚类方法得到说话人模型的初始化参数 λ_0.

Step 3 根据初始化参数 λ_0,利用 EM 算法计算得到每个说话人混合度 $M=128$ 的 GMM 模型, $G_L=\{\lambda_1,\cdots,\lambda_i,\cdots,\lambda_N\}$, λ_i 为第 i 个测试者 GMM 模型参数.

输出:G_L.

跨语种说话人确认算法:

输入:某一测试者汉语语音段 s_t, N 个说话人的 GMM 模型 G_L,阈值 $Threshold$($Threshold=20$).

Step 1　计算和提取测试者汉语语音的 17 维融合特征参数 F_t；

Step 2　计算 F_t 与假定的说话人 GMM 模型参数 λ_i 似然概率得分 $P(F_t \mid \lambda_i)$，并与阈值 *Threshold* 比较，大于阈值时接受测试者为说话人，否则拒绝. $P(F_t \mid \lambda_i) = \prod_{t=1}^{T} P(f_t \mid \lambda_i)$.

输出：确认结果.

3　实验结果及分析

本文的跨语种说话人语料库采集自 40 位测试者的语音，测试者分为 4 组，每组 10 人，每组采集{英语，汉语}、{朝鲜语，汉语}、{日语，汉语}、{蒙古语，汉语}等 4 种语音对，不限定语音文本内容，每人采集 12 min 的语音(两个语种时长各为 6 min). 学习算法随机选取 1 min 语音作为训练样本，剩余语音根据各个实验的需求分成若干时间长度的小段用于测试.

3.1　单语种说话人确认实验

对本文的融合特征与文献[9]中的 40 维融合特征进行了准确率和平均判别时间的对比实验，实验结果见表 1. 从表 1 中可以看出：本文方法仅提取包含浊音部分的语音段特征参数包含更多的个人语音信息，从而使准确率更高. 此外，由于本文去除了无用的静音段和清音段降低了特征参数的维度，从而减少了计算复杂度，因此平均判别时间上优于文献[9]的特征参数.

表 1　2 种特征参数下准确率和平均判别时间的对比

特征参数	准确率/%	平均判别时间/s
文献[9]的特征参数	85.3	8
本文的特征参数	90.5	5

3.2　跨语种说话人确认实验

本实验将测试者的英语、朝鲜语、日语和蒙古语等 4 个语种用于学习生成说话人模型，采用汉语作为测试语种进行说话人确认. 测试语音划分为时长 6 s 的语音片段，得到每组汉语语音段数为 600. 实验结果见表 2.

表 2　跨语种说话人确认的准确率

训练语种	测试语音段数	正确确认的语音段数	准确率/%
英语	600	518	86.3
朝鲜语	600	530	88.3
日语	600	536	89.3
蒙古语	600	527	87.8
平均准确率/%			88.0

表 2 表明，跨语种说话人确认的准确率低于单语种说话人. 其原因是汉语发音有声调，而英、朝、日、蒙[15-18] 4 种语言没有. 另外这 4 种语言中：①英语识别率最低. 其原因是英语采用的是口腔后部发音体系，即主要利用口腔后部的微小动作发音，而汉语是口腔前部发音体系，并且英语发音时发音器官比较紧张，而汉语发音较松弛，这些差异直接影响了说话人的声道发音特性相关的基音周期参数. ②朝鲜语、日语以及蒙古语的准确率接近. 其原因是汉、朝、日、蒙同属东亚地区的语言，使它们呈现出亲缘化的特征，换句话说朝、日、蒙 3 种语言与汉语较为相似.

本实验与文献[9]进行的跨语种说话人确认的平均判别时间的对比实验结果见表 3. 从表 3 可以看出，在识别准确率方面本文方法更占优势，这是因为本文方法减少了语种因素带来的影响. 另外，在跨语种说话人确认中由于测试语音与训练语音特征在空间分布的质心距相对增大，致使测试阶段的平均判别时间会比单语种的长，但是本文方法的平均判别时间仍低于文献[9]的方法.

表3 本文方法与文献[9]方法的跨语种平均准确率和平均判别时间的对比

测定方法	平均准确率/%	平均判别时间/s
文献[9]方法	79.5	10
本文方法	88.0	6

4 结论

本文提出了一种提取浊音段基音周期和加权的MFCC参数的融合特征的跨语种的说话人确认方法,实验结果证明了该方法的有效性和鲁棒性.与文献[9]的方法相比,本文方法在说话人确认的准确性和判别时间上都优于文献[9]的方法,由此表明本文方法不仅适用于单语种,也适用于跨语种的说话人确认.下一步研究工作中需要考虑提高说话人模型的准确性,可以与GMM-UBM、JFA、*i*-vector模型等进行对比研究.

参考文献:

[1] 郑方.声纹识别技术及其应用现状[J].信息安全研究,2016,2(1):44-57.
[2] 陈强.基于GMM的说话人识别系统研究与实现[D].武汉:武汉理工大学,2010:6-7.
[3] Sarkar S, Rao K S, Nandi D. Multilingual speaker recognition on Indian languages[C]//2013 Annual IEEE India Conference (INDICON). Mumbai, India, 2013:1-5.
[4] Bhattacharjee U, Sarmah K. A multilingual speech database for speaker recognition[C]//IEEE International Conference on Signal Processing, Computing and Control. Hong Kong, China, 2012:1-5.
[5] Ma B, Meng H. English-Chinese bilingual text-independent speaker verification[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal, Canada, 2004:V-293-6.
[6] Lu L, Dong Y, Zhao X. The effect of language factors for robust speaker recognition[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Taipei, Taiwan, 2009:4217-4220.
[7] 骆启帆.基于声门信息的说话人确认方法研究[D].杭州:杭州电子科技大学,2014:16-17.
[8] 房安栋.复杂背景下声纹特征提取与识别[D].长沙:中南林业科技大学,2014:35-45.
[9] Zhang Xuefeng, Dong Yuan. Insight into the role of pitch information in text-independent speaker recognition[C]//第八届全国人机语音通讯学术会议.北京,2005:214-217.
[10] Kim S, Eriksson T, Kang H G. A pitch synchronous feature extraction method for speaker recognition[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal, Canada, 2004:I-405-8.
[11] 王义元,赵黎明.基于小波变换和Teager能量算子浊音段提取[J].控制工程,2004,11(S2):99-101.
[12] Derrien T, Johnson R, Bussotti G. Wavelet speech enhancement based on the Teager energy operator[J]. IEEE Signal Processing Letters, 2001,8(1):10-12.
[13] Teager H. Some observations on oral air flow during phonation[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1980,28(5):599-601.
[14] 刘士.基于GMM的声纹识别技术的研究[D].成都:电子科技大学,2012:34-35.
[15] 曹仁松.汉语声调特点对英语语调学习的负迁移[J].大连海事大学学报(社会科学版),2008,7(3):189-191.
[16] 徐世荣.抓住声调教学这一环—突破朝鲜族学汉语的难点[J].汉语学习,1980(5):1-5.
[17] 张丽莉.日本初学者上声习得偏误分析及解决策略[D].大连:辽宁师范大学,2014:1-4.
[18] 侯红霞.蒙古国UB升日中学初级学生汉语声韵调习得偏误分析与应对策略[D].西安:西北大学,2014:34-37.