广西科学 Guangxi Sciences 2007,14(4):423～425

一种新的汉语方言辨识特征*

A New Features of Chinese Dialects Indentification

顾明亮[1,2]

GU Ming-liang[1,2]

(1.徐州师范大学物理与电子工程学院,江苏徐州 221116;2.徐州师范大学语言研究所,江苏徐州 221116)

(1. School of Physics and Electronic Engineering, Xuzhou Normal University, Xuzhou, Jiangsu 221116, China; 2. Linguistic Institution, Xuzhou Normal University, Xuzhou, Jiangsu, 221116, China)

摘要:将声学特征与韵律特征相结合,提出一种新的混合区间特征,并将该特征和常见的美尔倒谱系数(MFCC)特征与线性预测倒谱系数(LPCC)特征进行对比,通过符号化语言辨识方法对北方方言、吴方言、粤方言和闽方言进行辨识,以验证混合区间特征的有效性。结果表明,混合区间特征比MFCC特征和LPCC特征具有更好的方言辨识效果,对4种汉语方言15s语音片段的方言辨识率可以达到92%。4种方言中,混合区间特征对闽方言和粤方言的识别率最高,分别达到了96%和95%。

关键词:语音辨识 汉语方言 韵律特征 声学特征 GMM符号化器

中图法分类号:TP391.4 **文献标识码**:A **文章编号**:1005-9164(2007)04-0423-03

Abstract:Combining acoustic features with prosodic features, this paper presents a new hybrid block feature. In order to test the efficiency of the new feature, comparative experiments are done on the speech database consisting of North, WU, YUE and MIN dialects. The experimental results show that the new feature can performs better than traditional MFCC and LPCC features. An average accuracy of 92% is achieved in four Chinese dialects with 15 seconds speech segments. And the identification accuracy of MIN and YUE dialects is best in four dialects. They are 96% and 95% respectively.

Key words:identification, Chinese dialects, prosodic features, acoustic feature, GMM tokenizer

汉语方言自动辨识是计算机判别输入语音所属方言区域的过程,该技术既可以作为方言语音识别和自动咨询服务系统的前端,也可以直接应用到银行、宾馆、旅游等服务行业,还可在公安刑事侦查中确定罪犯的籍贯。汉语方言自动辨识研究还处于起步阶段,我国台湾学者蔡伟和等[1]研究相对较早,最近,新加坡学者B. P. Lim[2]也有相关成果介绍。我国内地有语言辨识[3]和汉语普通话口音辨识研究[4]的报道,但是汉语方言自动辨识的报道相对较少,这与国内没有合适的方言语音库有关,也与工程技术人员与方言研究者交流不够有关。

方言辨识的研究[5]起源于语言辨识。目前语言辨识的主流技术有两种:一种是Zissman[6]等提出的平行音素识别加语言建模的方法(PPRLM),它的主要优点是识别率高,在美国标准局组织的多次测评中名列前茅,但该方法需要有标注过的语音数据做训练样本,运算复杂度高,处理时间长;另一种是最近正在热烈讨论的符号化语言辨识方法[7],它利用高斯混合模型(GMM)进行音素的符号化处理,其优点是不需要标注的训练语音样本,运算复杂度低,处理时间短,存在的主要问题是识别率比PPRLM方法略低。

按照模式识别的原理,系统辨识效果的好坏,主要取决于特征的选取与分类器的设计。目前,各类语种识别所用的特征主要是声学特征或音联特征。但是

收稿日期:2007-03-28

修回日期:2007-07-05

作者简介:顾明亮(1963-),男,副教授,主要从事数字语音信号处理、神经网络理论与应用、模式识别、机器学习等研究。

* 江苏省"十五"社科基金项目(K3-013)和江苏省高校自然科学基金项目(99KJB510002)资助。

单独使用一种特征其效果往往受到一定限制，汉语方言是有调语言，音素间的差别相对较小，而韵律特征则对区分不同方言具有重要的影响。因此，本文将声学特征与韵律特征相结合，提出一种新的汉语方言辨识混合区间特征，并将该特征和常见的美尔倒谱系数(MFCC)特征与线性预测倒谱系数(LPCC)特征进行对比，通过符号化语言辨识方法对北方方言、吴方言、粤方言和闽方言4种方言进行辨识，以验证混合区间特征的有效性。

1 混合区间特征的原理和方法

混合区间特征的原理如图1所示。图1中，预处理模块与后续处理模块有关，主要包括：预加重，分帧，有声/无声检测等。MFCC 特征是语音识别中应用最多的特征，特征提取时首先将频谱转化为基于美尔(Mel)频标的非线性频谱，然后再转换到倒谱域上[8]。基频提取算法采用时域法来计算各帧的基音频率[9]。

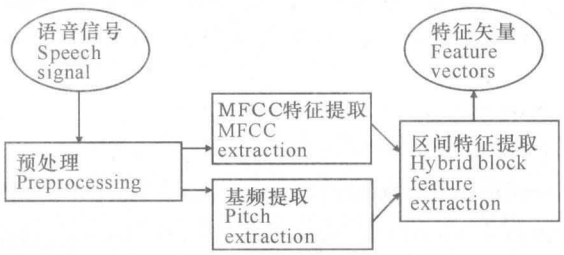

图1 混合区间特征提取

Fig. 1 A framework for hybrid block feature extraction

混合区间特征的提取方法如下。

(1)计算各帧的 MFCC 及基频(F0)，设第 t 帧的 MFCC 和基频为：$c_i(t)i = 1,2,\cdots,N$，其中，第 N 个系数代表基频 F0；

(2)利用 MFCC 与 F0计算差分系数：$\Delta c_i(t) = c_i(t+1) - c_i(t-1), i = 1,2,\cdots,N$；

(3)计算一阶区间差分倒谱：$\Delta e_i(t) = c_i(t+P+1) - c_i(t+P-1)$，其中，$P = 3; i = 1,2,\cdots,N$；

(4)计算二阶区间差分倒谱：$\Delta d_i(t) = c_i(t+2P+1) - c_i(t+2P-1)$，其中，$P = 3; i = 1,2,\cdots,N$；

(5)由此可得到第 t 帧的总的特征矢量为：$[\Delta c_1, \cdots, \Delta c_N; \Delta e_1, \cdots, \Delta e_N, \Delta d_1, \cdots, \Delta d_N]$，特征矢量的维数为 $3N$。本实验中 $N = 9$。

2 实验比较与分析

在文献[10]的方言辨识系统基础上，改变语音特征部分和影响系统辨识性能的参数，进行汉语方言辨识实验。

2.1 语音库

语音库选择4种比较有代表性的方言：北方方言、吴方言、粤方言和闽方言。每种方言说话人为10～12人，并按男女比例1:1选取。语音库分为3个部分：训练集、测试集和开发集。训练集用于训练 GMM 符号化器和语言模型，测试集用于测试整个系统的性能，开发集用于系统性能提升比如训练后端分类器。训练集中每种方言各有一个约60min 的训练语料，测试集和开发集都是5s、10s 和15s 的语音段的集合，测试集中每种方言各有60段3种时长的测试语音，同样地开发集中每种方言也各有60段3种时长的语音。以上3个语音集语音互不交叉重叠。

2.2 系统参数对辨识结果的影响

利用4种方言构成的语音数据库，对 GMM 阶数分别为16、32、64和128阶，测试语音长度为15s 时的系统性能进行的实验结果如图2所示。

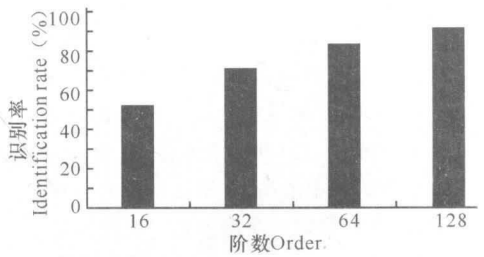

图2 GMM 阶数对系统性能的影响

Fig. 2 The relationship between the order of GMM and system performance

由图2可见，随着 GMM 阶数的提高，系统性能也随之有所提升。这是因为 GMM 阶数增加，模型的刻画能力越强。

采用神经网络后端分类器和一般语种辨识中常采用的高斯后端分类器测试分类器对系统性能的影响实验结果见表1。系统的 GMM 阶数是128阶，GMM 高斯符号化器个数从1增加到4。

表1 不同后端分类器对系统性能的影响比较

Table 1 Compared the system performance under different backend classifiers

符号化器 Tokenizer(个)	识别率 Identification rate(%)	
	神经网络分类器 ANN	高斯分类器 GMM
1	56	43
2	72	56
3	84	63
4	92	66

从表1可以看出，与使用高斯后端分类器相比，系统使用神经网络后端分类器时，系统的辨识率得到大幅提升，特别是对于并行系统辨识率更是提高了26%。

2.3 不同特征对系统性能的影响

选择 GMM 阶数为128，GMM 符号化器个数为

4，分类器采用神经网络后端分类器作为参数，将该特征和常见的美尔倒谱系数(MFCC)特征与线性预测倒谱系数(LPCC)特征进行对比，通过符号化语言辨识方法对北方方言、吴方言、粤方言和闽方言进行辨识性能实验的结果见表2。

表2 不同特征下的系统识别结果

Table 2 Compared identification results under different features

语音段 Segment of sound (s)	识别率 Identification rate(%)				
	12维 MFCC 12 MFCC	12维 MFCC +12维一阶差分 12 MFCC + 12 ΔMFCC	12维 LPCC 12 LPCC	12维 LPCC +12维一阶差分 12 LPCC + 12ΔLPCC	27维混合区间特征 27 HBF
5	72	75	73	76	85
10	81	83	80	82	88
15	86	87	83	85	92

表2结果表明，无论对短时还是长时语音，混合区间特征均有最好的识别性能，对4种汉语方言15s语言片段的方言辨识率达到92%。MFCC特征对长时语音辨识效果好，LPCC特征对短时语音更有效，这可能与LPCC有较强的短时预测性有关。

系统在测试15s语音段时，混合区间特征对4种方言的辨识率分别为：北方方言87%，吴方言90%，粤方言95%，闽方言96%，闽、粤、吴方言较北方方言具有更高的识别率，说明方言声调对于声调丰富的方言具有更好的辨识效果。

3 结束语

本文在总结和分析语种辨识特征的基础上，结合汉语方言的语音特点，提出了基于声学特征与韵律特征相结合的混合区间特征，该特征既保留了高效的MFCC特征，又充分利用了对汉语辨识有重要作用的声调特征。实验表明，混合区间特征对于区分4种不同方言效果明显，其中闽方言和粤方言识别更好，北方方言相对差一些，这与北方方言缺少调类变化有关。

致谢：

感谢中国科学院自动化所黄泰翼研究员和清华大学电子工程系王作英教授在系统设计中的热情鼓励和指导，感谢江苏省语言科学与神经认知工程重点实验室主任杨亦鸣教授在方言数据库建设过程中的热情帮助。

参考文献：

[1] TSAI WUEI HE, CHANG WEN W HEI. Discriminative training of gaussian mixture bigram models with application to chinese dialect identification [J]. Speech Communication, 2002, 36: 317-326.

[2] LIM BOON PANG, LI HAIZHOU, MA BIN. Using local & global phonotactic features in chinese dialect identification [C]. IEEE International Conference on Acoustics Speech and Signal Processing, 2005: 577-580.

[3] 屈丹，王炳锡，魏鑫. 基于GMM-UBM模型的语言辨识研究[J]. 信号处理，2003，19(1)：85-88.

[4] CHEN TAO, HUANG CHAO, CHANG ERIC, et al. Automatic accent identification using gaussian mixture models: proceedings IEEE automatic speech recognition and understanding workshop[C]. Italy, 2001.

[5] MUTHUSAMY Y K, BARNARD E, COLE R A. Reviewing automatic language identification [J]. IEEE Signal Processing Magazine, 1994(10): 33-41.

[6] ZISSMAN M A. Comparison of four approaches to automatic language identification of telephone speech[J]. IEEE Trans Speech and Audio Pro, 1996, 4(1): 31-34.

[7] TORRES-CARRASQUILLO P A, REYNOLDS D A, DELLER J R J R. Language identification using gaussian mixture model tokenization [C]. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP '02), 2002: 13-17.

[8] DAVIS S B, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Transaction on Acoustics Speech and Signal Processing, 1980, 25(4): 357-366.

[9] SHIMAMURA T, KOBAYASHI H. Weighted autocorrelation for pitch extraction of noisy speech [J]. IEEE Trans Speech Audio Processing, 2001, 9(7): 727-730.

[10] 顾明亮，沈兆勇. 基于语音配列的汉语方言自动辨识[J]. 中文信息学报，2006，20(5)：77-82.

(责任编辑：韦廷宗 邓大玉)