基于滑动倒谱的自动语言辨识

王洪海,刘 刚,郭 军
(北京邮电大学 信息工程学院,北京 100876)

摘 要:滑动差分倒谱在自动语言辨识的研究中获得了广泛的应用.但是滑动差分倒谱并没有利用语音信号的静态倒谱信息,在方言辨识中的研究表明静态倒谱比差分倒谱含有更多的特征信息.为此,提出了滑动倒谱(SC)的概念,并与滑动差分倒谱特征矢量进行了对比研究.首先利用开发集的语音考察了滑动差分倒谱和滑动倒谱的控制参数在不同取值的情况下对识别性能的影响,利用爬山法确定了这 2类特征矢量达到局部最优控制参数组合的路径,然后利用测试集的数据对优化后的 2类特征矢量建立的模型进行了闭集辨识和开集辨识.2种情况下的测试结果都表明滑动倒谱的性能优于滑动差分倒谱.并且这 2种参数还具有特征互补性,将它们进行决策级数据融合可以进一步提高系统的性能.

关键词:自动语言辨识;滑动倒谱;滑动差分倒谱;高斯混合模型
中图分类号: TP391.42 **文献标识码**: A **文章编号**: 1673-4785(2008)04-0336-06

Automatic language identification using shifted cepstra

WANG Hong-hai, LIU Gang, GUO Jun
(Information Engineering College, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Shifted delta cepstra have been widely used in automatic language identification, but only delta cepstrum information is employed. Research on accent identification revealed that detailed cepstrum is more informative than delta cepstrum. So shifted cepstrum was proposed and comparative study was conducted between these two cepstra. Effects of their control parameters on recognition performance were investigated with speech data in the development set. The best paths of these two vectors to reach a locally optimal control parameter combination were determined with the hill-climbing method. Comparative tests performed with speech data both in the closed test set and open test set demonstrated that shifted cepstra is superior to shifted delta cepstra. In addition, they are mutually complementary and data fusion at the decision level could further improve the performance of the system.

Keywords: automatic language identification; shifted cepstra; shifted delta cepstra; Gaussian mixture model

基于声学特征的方法是自动语言辨识(automatic language identification, ALID)研究中经常采用的一种方法[1].它直接利用不同语言之间的频谱或倒谱差异作为语言识别的依据,因而具有计算复杂度低、可移植性好及不需要音素标注的训练语料等优点.实验表明,基于声学特征的高斯混合模型(Gaussian mixture model, GMM)对 12种语言的识别完全可以做到实时处理,而基于并行的音素识别结合语言模型(parallel phoneme recognition followed by language modeling, PPRLM)的系统则需要 14倍的实时处理时间[2].尤其是随着滑动差分倒谱(shifted delta cepstra, SDC)参数在 ALID研究中的成功应用,使得基于声学特征的研究方法获得了突破性的进展[3-4].在 2003年美国国家标准与技术协会(National Institute of Standards and Technology, NIST)组织的自动语言辨识系统评测中,采用 SDC参数的

收稿日期: 2007-06-28.
基金项目:"十一五"国家 863计划重点项目课题(2006AA010102)
通信作者:王洪海. E-mail: greegrassw@sina.com.

GMM方法获得了比 PPRLM方法更好的识别效果[2],改变了人们长期以来的观点[5]. 如今, SDC参数在 ALID研究中获得了广泛的应用[6~9].

SDC是差分倒谱系数的扩展,它同时考虑了前后帧差分倒谱的影响,具有融合长时序信息特征的能力. 但是, SDC只利用了差分倒谱信息,即语音信号的动态特性,并没有考虑语音信号的静态特性,即频谱/倒频谱信息. 而 WU Tingyao等人在方言识别中的研究表明,详细刻画的频谱/倒频谱比差分倒谱包含更多的信息[10]. Matejka Pavel等人在语言辨识的研究中将美尔倒频谱系数 (Mel frequency cepstral coefficients, MFCC)与 SDC系数结合在一起获得了比单独的 SDC更好的识别效果[7]. 因此,本文根据 SDC的思想提出了滑动倒谱 (shifted cepstra, SC)的概念,与目前研究中常用的 SDC特征矢量进行了对比研究.

1　滑动差分倒谱

滑动差分倒谱参数由若干块跨多帧语音的差分倒谱组成,这样在一个特征矢量内包含多帧语音的长时声学信息. 差分倒谱参数一般通过式 (1)计算:

$$\delta_j(t) = C_j(t+d) - C_j(t-d), \quad j = 0, 1, \cdots, N-1. \tag{1}$$

式中: $C_j(t)$是第 t帧语音中第 j个倒谱系数,每帧语音中包含 N个倒谱系数.

滑动差分倒谱通过串联 k块差分倒谱在一帧内对差分倒谱进行了扩展,其中每块差分倒谱向后滑动了 p帧,其表达式为

$$\begin{aligned} S(t) = [&\delta_0(t), \delta_1(t), \cdots, \delta_{N-1}(t), \delta_0(t+p), \\ &\delta_1(t+p), \cdots, \delta_{N-1}(t+p), \delta_0(t+(k-1)p), \\ &\delta_1(t+(k-1)p), \cdots, \delta_{N-1}(t+(k-1)p)]. \end{aligned} \tag{2}$$

这样,每帧内的差分倒谱系数由 N个扩展到了 kN个. SDC特征向量由 4个参数确定:每帧语音中包含的倒谱系数个数 N,计算差分倒谱的时移 d,差分倒谱块的滑动帧数 p和一个 SDC特征向量中包含的差分倒谱块的个数 k

Kohler等人的研究表明,不同的 N-d-p-k参数组合对系统识别性能的影响不同. 最佳的参数组合与所要识别的语言类型有关[4].

2　滑动倒谱

根据 SDC的思想可以直接在静态倒谱的基础上构建滑动倒谱 SC,即直接在每一帧内串联 k块倒谱系数,其中每块倒谱向后滑动了 p帧,其表达式为

$$\begin{aligned} S(t) = [&C_0(t), C_1(t), \cdots, C_{N-1}(t), C_0(t+p), \\ &C_1(t+p), \cdots, C_{N-1}(t+p), C_0(t+(k-1)p), \\ &C_1(t+(k-1)p), \cdots, \\ &C_{N-1}(t+(k-1)p)]. \end{aligned} \tag{3}$$

式中: $C_j(t)$是第 t帧语音中第 j个倒谱系数. 这样,每帧内的倒谱系数由 N个扩展到了 kN个. SC特征向量由 3个参数确定:每帧语音中包含的倒谱系数个数 N,倒谱块的滑动帧数 p和一个 SC特征向量中包含的差分倒谱块的个数 k

从滑动倒谱的构成可以看出,它与滑动差分倒谱一样,可以在一个特征向量内融入比较长的时序信息,因而它能够刻画长时间间隔的过渡期信息特征. 听觉实验研究表明,人类的听觉特性对语音频谱的过渡信息非常敏感,虽然差分倒谱参数可以描述 50～100 ms时间间隔的过渡信息特征,但是它却无法刻画更长时间间隔如 200～300 ms的长过渡期信息特征,而这种长过渡期信息对应着音素到音素、音节到音节的过渡信息. Furui曾认为,如何采用一种特征参数形式描述长过渡期的语音信息特征是一个有待解决的问题[11]. 而从 SDC和 SC特征向量的结构特性来分析,这 2种参数形式为解决这一问题提供了借鉴思路,因为它们都能够融合长时间间隔的信息特征. 至于需要这 2种参数形式刻画多长时间间隔的过渡期特征,可以结合具体的任务系统通过实验调整这 2种特征向量的控制参数组合来实现.

3　实验和分析

3.1　语音语料库

实验中所用的汉语语音来源于 863汉语普通话语料库,英语、日语、德语、法语、西班牙语、俄语和阿拉伯语等 7个语种的语音是从网络上采集的,每个

语种包含了多种内容体裁. 整个语料库的语音被分成训练集、开发集和测试集 3部分. 训练集包括汉、英、日、德、法、西 6个语种,每个语种包括 36～38个说话人,每个说话人的语音片段为 30～60 s,每种语言大约有 20 min的训练语料. 开发集也只包括汉、英、日、德、法、西 6个语种,每个语种包含 5名男性和 5名女性的语音,每人有 50个平均时长为 4.5 s的语音片段. 测试集包括闭集和开集 2个集合. 闭集包括汉、英、日、德、法、西 6个语种,与训练集中的语种完全相同,而开集则在闭集的基础上增加了俄语和阿拉伯语. 测试集中,每个语种包括 10名男性和 10名女性、每人有 50个语音片段,每个测试语音片段的平均长度为 4.5 s训练集、开发集和测试集中的说话人没有交叉. 关于语料库的详细介绍请参见文献[12].

3.2 对开发集的实验

对于开发集的实验主要是考察不同的控制参数组合对 SDC和 SC特征向量的性能的影响,利用爬山法确定这 2类特征矢量达到局部最优识别效果时的控制参数组合,并对这 2类特征向量采用加权系统数 α进行数据融合.

实验中,输入的语音经 16 kHz取样 16 bit量化后进行预加重,用帧长为 25 ms、帧移为 10 ms的汉明窗分帧,计算 13维的 RASTA-PLP参数(包括 0阶的能量系数). 然后,取 N-p-k为 13-3-3构建 SC特征矢量,利用期望最大算法为每种语言建立 GMM模型.

与此同时,在经 RASTA滤波的感知线性预测(RASTA-pereteptual linear prediction, RASTA-PLP)参数的基础上计算差分倒谱,然后取 N-d-p-k为 13-1-3-3构建 SDC特征矢量,并为每种语言建立 GMM模型. 所有 GMM模型的混合分量数目都为 128. 这样,对应于 SC和 SDC控制参数组合的一组初始值分别建立起了系统的初始模型,对于开发集中的语音进行测试的结果见表 1.

表 1 初始模型的测试结果

Table 1 Test results of original model

特征参数	误识率/%
RASTA-PLP-SDC	2.37
RASTA-PLP-SC	1.57

以下就从系统的初始模型(SDC和 SC的控制参数组合分别为 13-1-3-3和 13-3-3)出发,依次调整特征向量的控制参数,考察它们对性能的影响.

3.2.1 参数 N对性能的影响

首先,保持 2类特征向量的其他控制参数不变,只调整参数 N的取值,得到测试结果如表 2所示. 从表 2中可以看出,对于普遍使用的 13维的 RASTA-PLP倒谱系数,其 SDC和 SC特征矢量并没有表现出最好的识别性能,而是在阶数比较少的情况下(分别是 9维和 7维)SDC和 SC参数获得了比较好的识别效果. 这说明,对于 SC和 SDC,比较少的系数已经包含了充分的识别信息,信息冗余反而会造成识别性能下降.

表 2 参数 N对性能的影响

Table 2 Effect of N on performance

特征参数	误识率/%	特征参数	误识率/%
SDC(13-1-3-3)	2.37	SC(13-3-3)	1.57
SDC(10-1-3-3)	2.17	SC(10-3-3)	1.73
SDC(9-1-3-3)	2.07	SC(9-3-3)	1.53
SDC(8-1-3-3)	2.57	SC(7-3-3)	1.50
SDC(7-1-3-3)	2.33	SC(6-3-3)	2.43

3.2.2 参数 k对性能的影响

对于 SC特征矢量,使控制参数在组合 7-3-3的基础上调整 k的取值,而对于 SDC特征矢量,以前的研究中[2,7-9]得到的 N的优化数值为 7,所以本实验中取 SDC的控制参数分别为 7-1-3-3和 9-1-3-3为初值,然后调整 k的取值,利用开发集中的语音进行测试得到了表 3中所列的结果.

表 3 参数 k对性能的影响

Table 3 Effect of k on performance

特征参数	误识率/%	特征参数	误识率/%
SDC(9-1-3-5)	1.77	SC(7-3-3)	1.50
SDC(7-1-3-5)	1.53	SC(7-3-5)	1.07
SDC(7-1-3-6)	1.47	SC(7-3-6)	1.13
SDC(7-1-3-7)	1.67	SC(7-3-7)	1.30

从表 3中可以看出,对于滑动倒谱 SC矢量,当串联倒谱块的数目为 5时表现出了最好的性能,当 k继续增加时,系统的识别率略有降低. 而对于滑动差分倒谱 SDC矢量,虽然在控制参数为 9-1-3-3时

的识别效果好于 7-1-3-3,但是,当 k增加到 5时,由 SDC(7-1-3-5)得到的改善效果明显好于 SDC(9-1-3-5).因此,接下来继续对 SDC(7-1-3-6)和 SDC(7-1-3-7)进行测试,并由此确认 SDC(7-1-3-6)可以达到局部最优的识别效果.

3.2.3　参数 p对性能的影响

p是相邻倒谱块的相对滑动帧数,它确定了进行信息融合的前后帧的时移.确定最佳的 p值可以说明前后哪些帧的参数具有最大的互补性.在 SDC和 SC特征矢量的控制参数分别取 7-1-3-6和 7-3-5的基础上调整 p的取值,测试结果如表 4所示.

表 4　参数 p对性能的影响

Table 4　Effect of p on performance

特征参数	误识率/%	特征参数	误识率/%
SDC(7-1-3-6)	1.47	SC(7-3-5)	1.07
SDC(7-1-2-6)	1.40	SC(7-2-5)	1.03
SDC(7-1-1-6)	1.90	SC(7-1-5)	1.17

从表 4中可以看出,滑动 2帧的倒谱块参数具有最大的互补性,这些互补性的信息融合在一个 SC特征向量内可以达到比较好的识别效果.与 SC特征向量一样,SDC特征向量也需要融合滑动 2帧的差分倒谱块才能达到比较好的效果.

3.2.4　参数 d对 SDC矢量性能的影响

SDC矢量比 SC矢量多了一个控制参数 d,它是计算差分倒谱的时移.当 d值变化时对 SDC(7-X-2-6)矢量性能的影响如表 5所示.

表 5　参数 d对 SDC性能的影响

Table 5　Effect of d on perfomance of SDC

特征参数	误识率/%
SDC(7-1-2-6)	1.40
SDC(7-2-2-6)	2.33

从表 5可以看出,按照前后帧的时移间隔计算差分倒谱可以使 SDC特征向量获得比较好的性能.

从以上实验可以看出,调整滑动倒谱 SC和滑动差分倒谱 SDC的控制参数可以使系统的识别率得到明显的改善.对于特定的语音语料库和识别任务,SDC矢量和 SC矢量应当各自存在一个最优的参数组合,使系统的识别性能达到最佳.但是,最优的控制参数需要长时间的搜索才能确定.一般通过爬山法可以比较快捷地得到一个局部最优的参数组合.通过 3.2.2节的实验可知,局部最优的参数未必是全局最优的.有时需要利用经验知识对搜索方向进行调整.图 1给出了利用爬山法和经验知识进行搜索确定的控制参数优化路径,并标出了对应的控制参数.

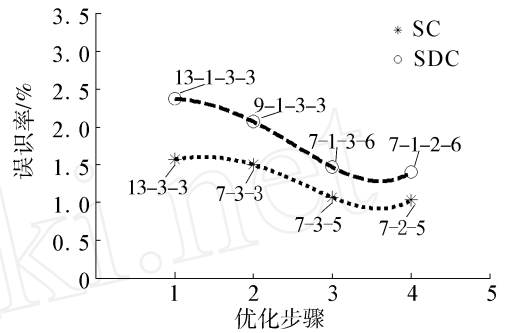


图 1　SC和 SDC控制参数优化的路径

Fig 1　Optimizing path of control parameters of SC and SDC

3.2.5　SC与 SDC矢量的数据融合

利用以上实验中性能达到局部最优的 SC(7-2-5)参数所建立的模型作为一个子系统(标注为 Sub1),然后与采用 SDC(7-1-2-6)参数建立的子系统(标注为 Sub2)进行决策级数据融合,2个子系统使用的分类器都是混合分量数为 128的 GMM.融合方式采用线性加权组合,即:

$$S = S_{\mathrm{Sub1}} + \alpha \times S_{\mathrm{Sub2}}. \qquad (4)$$

式中:S_{Sub1}和 S_{Sub2}分别代表 2个子系统的得分,S为数据融合之后系统的总得分.式(4)表示首先固定 Sub1子系统的加权系数为 1,然后调整 Sub2子系统的加权系数 α,使融合后的识别效果达到全局或局部最优.加权系数 α采用搜索算法确定,即从 $\alpha = 1$开始,按 0.1的步长增加或降低 α的取值,使系统的识别率逐步增加,直到达到一个局部最优的结果.表 6给出了最终确定的加权系数及对应的测试结果.

表 6　决策级数据融合

Table 6　Data fusion on decision level

特征参数	加权系数 α	误识率/%
SC(7-2-5)	…	1.03
SDC(7-1-2-6)	…	1.40
数据融合	0.1	0.97

3.3 对测试集的实验

对测试集的实验分为闭集辨识和开集辨识. 对于开集辨识,要求系统首先判决被测语言片段的语种是否属于闭集中的注册成员. 因此,开集辨识比闭集辨识多了一个确认过程,其正确识别率将有所降低,但与实际情况更为接近.

3.3.1 闭集辨识

根据 3.2节得到优化结果,分别选取最优的特征参数 SC(7-2-5)和 SDC(7-1-2-6)建立模型对测试集闭集中的数据进行测试,然后再利用优化的加权系数将 2类模型进行融合,得到了表 7所列出的测试结果.

表 7 对测试集的闭集辨识结果

Table 7 Identification results for the closed test set

特征参数	加权系数 α	误识率 /%
SC(7-2-5)	…	2.05
SDC(7-1-2-6)	…	2.32
数据融合	0.1	1.97

从表 7可以看出,滑动倒谱的性能优于滑动差分倒谱,将滑动倒谱与滑动差分倒谱进行数据融合可以进一步提高系统的识别率.

3.3.2 开集辨识

在开集辨识中,系统首先根据设定的阈值对被测语言片段的语种是否属于闭集做出判决,此时使用拒识率 E_{FJ}和误识率 E_{FA} 2个参量来表征系统的性能. 调节判决阈值的大小可以得到拒识率和误识率相等时的等错误率 (equal error rate, EER).

根据优化结果,分别选取最优的特征参数 SC(7-2-5)和 SDC(7-1-2-6)建立模型对测试集开集中的语言片段是否属于闭集中的语种进行表决,在不同的判决阈值条件下得到不同的拒识率和误识率. 通过调节判决阈值的大小得到最后的 EER,结果见表 8.

表 8 对测试集开集的确认结果

Table 8 Verification results for the open test set

特征参数	加权系数 α	EER/%
SC(7-2-5)	…	6.83
SDC(7-1-2-6)	…	7.35
数据融合	0.1	6.28

然后,再利用优化的加权系数将 2类模型进行融合,重新设定判决阈值进行表决,得到了数据融合后的 EER,如表 8所示.

对于初步确认语种属于闭集的语音片段进行进一步的识别,以确定其具体的语言种类,得到表 9所示的识别结果.

表 9 对测试集的开集辨识结果

Table 9 Identification results for the open test set

特征参数	加权系数 α	误识率 /%
SC(7-2-5)	…	8.33
SDC(7-1-2-6)	…	8.98
数据融合	0.1	7.85

从表 8和表 9中可以看出,对于开集辨识的语种确认过程和识别过程,使用滑动倒谱的效果也好于滑动差分倒谱. 并且这 2种参数也具有特征互补性,将它们进行数据融合可以进一步改善系统的识别效果.

4 结束语

从对比实验可以看出,无论对于闭集辨识的测试还是开集辨识的测试, SC参数比 SDC参数都表现出了更优越的性能,并且 SC参数不需要计算差分倒谱,计算量比 SDC参数小,因此,这种参数形式值得在今后的研究中进一步关注. 利用爬山法可以方便快捷地对 SC和 SDC的控制参数进行优化. 但是,这种方法得到的往往是局部最优结果,并且需要利用经验知识对搜索方向进行调整. 以前的研究[2,7-9]指出, SDC控制参数的优化组合为 7-1-3-7,而在本实验中得到的局部最优参数组合为 7-1-2-6,这说明最佳的控制参数组合与系统的识别任务及所使用的语音语料库密切相关. 另外, SDC和 SC特征向量中不同的控制参数组合最终反映了特征信息的时序长度和向量中内嵌特征块的间隔,而这 2项又同时受到帧长和帧移的影响. 因此,帧长和帧移变化时,最优的控制参数组合也可能受到影响,最终系统的性能也会有所变化. 而在以前的研究中还没有关于最优的控制参数组合随帧长和帧移变化的讨论,这种变化关系的确定需要进一步研究.

参考文献:

[1]王洪海, 刘 刚, 郭 军. 自动语言辨识研究方法及发展概述[J]. 电脑与信息技术, 2007, 16(2): 37-39.

WANG Honghai, LIU Gang, GUO Jun. Overview of approaches to automatic language identification and recent development[J]. Computer and Information Technology, 2007, 16(2): 37-39.

[2] SINGER E, TORRES C, GLEASON T P, et al. Acoustic phonetic and discriminative approaches to automatic language recognition[C]//Proc of Eurospeech. Geneva, 2003: 1345-1348.

[3] TORRES-CARRASQUILLO P A, SINGER E, KOHLER M A, et al. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features [C]// Proc of ICSLP. Denver, USA, 2002: 89-92.

[4] KOHLER M A. Language identification using shifted delta cepstra[C]// Proc of Midwest Symposium on Circuits and Systems [S.l], 2002: 69-72.

[5] ZISSMAN M A. Comparison of four approaches to automatic language identification of telephone speech[J]. IEEE Trans on Speech and Audio Processing, 1996, 4(1): 31-44.

[6] BO Yin, ELFATHAMBY A, FANG Chen. Combining cepstral and prosodic features in language identification[C]// Proc of 18th International Conference on Pattern Recognition. Hongkong, 2006: 254-257.

[7] CAMPBELL l W, GLEASON T, NAVRATIL l J, et al. Advanced language recognition using cepstra and phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation[C]// Proc of Odyssey: The Speaker and Language Recognition Workshop. San Juan, Ruerto Rico, 2006: 1-8.

[8] BURGET L, MATEJKA P, CERNOCKY J. Discriminative training techniques for acoustic language identification [C]// Proc of ICASSP. [S.l], 2006: 209-212.

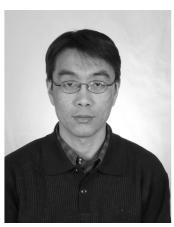
[9] MATEJKA P, BURGET L, SCHWARZ P, et al. Brno university of technology system for NIST 2005 language recognition evaluation[C]// Proc of Odyssey: The Speaker and Language Recognition Workshop. San Juan, Ruerto Rico, 2006: 57-64.

[10] WU Tingyao, COMPERNOLLE D V, DUCHATEAU J, et al. Spectral change representation and feature selection for accent Identification tasks[C]// Proc of the Workshop on Modeling for the Identification of Languages. Paris, 2004: 57-61.

[11] FURUI S. Recent advances in speaker recognition[C]// Proc of the First International Conference on Audio- and Video-based Biometric Person Authentication. [S.l], 1997: 237-252.

[12] 王洪海. 基于声学特征的自动语言辨识研究[D]. 北京:北京邮电大学, 2007.
WANG Honghai. Acoustic-based research on automatic language Iientification[D]. Beijing: Beijing University of Posts and Telecommunications, 2007.

作者简介:



王洪海,男,1970年生,高级工程师,主要研究方向为自动语言辨识,发表学术论文近 10篇.

刘　刚,男,1973年生,副教授,主要研究方向为语音识别、文字识别、语音合成等.

郭　军,男,1959年生,教授,博士生导师,北京市中高级职称评审委员会计算机组副组长,主要研究方向为模式识别、网络控制与管理等. 主持开发的基于整形变换的手写汉字识别方法在对日本国家标准汉字数据库 ETL9的测试中获最高识别率,在 1995年全国评测中获得识别率第 1名.