一种面向汉语语音识别的口形形状识别方法*

钟 晓[1] 周昌乐[2] 俞瑞钊[1]

[1](浙江大学计算机系智能软件实验室 杭州 310027)

[2](杭州大学计算机系视听实验室 杭州 310028)

摘要 借助汉语发音口形的生理特点,在音素识别这一水平上进行汉语语音的辅助识别,具体给出了一种口形形状识别和灰度的统计方法及其具体实现.实验结果基本与理论估算相吻合,对5个元音的口形区别正确率在80%以上,为语言的声波识别提供了一种有利的辅助手段.

关键词 汉语语音识别,口形特征提取,口语看话,统计模式识别.

中图法分类号 O235

人类进行语言交流时,特别是在环境噪声非常强的情况下,不仅使用声学言语信号来理解语言,通常还利用其他信息源,如口语看话(Lip-reading),面部表情(Facial-expression),手势(Hand-gesture)和身体语言(Body-language)等来识别语言.对于有听力障碍的聋哑人,口语看话更是一种高水平的语言交流手段.语言学的研究表明,仅靠看口形,聋哑人就可以理解一个句子的70%~80%的内容.口形及其动态变化在言语理解中的重要性由此可见.实际上,通过口形序列来识别语言音节,更有利于某些特殊领域的应用,如在隔音或远距离条件下的视觉"窃听"自动装置的研制、聋哑人的助听辅助工具的研制、言语的辅助识别和理解以及面部表情分析的临床应用等.

口形识别及其序列分析的研究是属于面孔自动识别和面部表情分析的研究范围.从能查阅到的资料来看,根据口形来进行语音识别的研究,国内尚无先例,国外则主要是一些初步的研究[1~7],其大体思想都是根据单幅口形图像,通过一定的形状特征提取,用于语言(特别是元音)的辅助识别或者作为语音识别的后处理,以便提高语音识别系统的识别率.就报道的结果看,尽管利用的信息是粗略简单的,但提高的语音识别率却相当可观.令人惋惜的是,这些研究由于只停留在单幅图像和简单特征的利用上,还存在许多遗憾.为此,我们对口形语音的变化进行了较为系统的研究,提出了一种基于形状拟合之上的统计模式识别方法,较为有效地解决了这一问题.

1 口形语音音素分析

语音口形,顾名思义就是人类语言交流中口形形态变化,它属于生理语音学的范畴.但由于一直以来人们在生理语音学上的研究主要是从听音、记音入手来研究的,也就是凭耳朵听辨语音进行分析研究,语音口形的研究一直处于未开发状态,但前人还是为我们留下了几个典型的语音口形[8~10]和一些系统的研究成果.

研究表明,人类在言语交流中形成的语音起码在4个方面直接或间接地与口形及其序列变化有关:

(1) 嘴唇的开口、动静、前撮等形状与音素音位有关;

(2) 口形大小、持续时间与音长和重音有关;

(3) 口形过渡形变与音渡有关;

(4) 序列口形组合与音节结构有关.

---

* 本文研究得到浙江省自然科学基金和北京大学视觉与听觉国家实验室基金资助.作者钟晓,女,1973年生,博士生,主要研究领域为智能软件.周昌乐,1959年生,博士,教授,主要研究领域为计算语言学,认知逻辑学,多媒体艺术.俞瑞钊,1937年生,教授,博士生导师,主要研究领域为智能软件.

本文通讯联系人:钟晓,杭州 310027,浙江大学 243 信箱

本文 1997-10-23 收到原稿,1998-01-23 收到修改稿

根据我们的分析，对于语音音素的口形特点，可以归纳成表 1 的情况，这也是我们进行口形音素识别的主要依据，其中音素是指元音和辅音.

表 1 音素口形描述表

音素口形代码	图示(手绘)	音素	总和
A		a	1
B		o	1
C		e	1
D		j q x i u ng l z c s	10
E		u	1
F		ü	1
G		er r	2
H		d t g k h	5
I		b p	2
J	双唇紧闭	m	1
K	上齿咬下唇	f	1
L	上齿咬下齿	zh ch sh	3
总计			29

2 口形特征选择和提取

我们知道，特征选择的目的在于对误差概率无不利影响时，减少所用特征的数目. 在实际应用中，特征的提取或选择基于已知的有限样本集 S 所供给的信息. 而在大多数视觉检验的应用方面，则是把图像数据与采用缺陷的特性或部件的量纲所进行的描述联系起来，即把图像数据映射为与检验有关的信息的表达式，而这种表达式能直接而容易地推导出来[11].

根据以上所述及语音学的知识、口形图像实例分析，我们选择圆唇度、展唇度、开口度和口形面积作为口形语音识别的基本特征. 其中各特征的确定和提取分述如下：

(1) 圆唇度：指唇的前撮度. 首先对所获取的口形图像进行预处理，得到边缘特征明显的新图像，再采用双目检测技术对其进行空间检测，提取圆唇度这一特征.

(2) 展唇度：指唇的宽度. 与上类似，先进行口形图像预处理，再对新图像进行扫描，提取展唇度这一特征.

(3) 开口度：指唇的垂直高度. 与上类似，对经过预处理的新图像进行列扫描，提取开口度这一特征.

(4) 口形面积：指唇内区域的面积. 经过实践检验，我们决定通过计算新图像边界内像素点的个数来提取口形面积这一特征.

这就是口形特征选择和提取的基本思想，详见第 4 节.

3 基于统计方法的口形识别原理

模式识别就是研究一种自动技术，依靠这种技术，机器将自动把待识别模式分配到各自的模式类中去. 统计模式识别的传统内容包括：几何分类法、概率分类法和聚类[12]. 我们经过具体分析和实践检验，决定选用聚类法进行口形识别.

聚类是一种无教师的分类法. 采用这种方法，从原理上讲，我们可以选用误差平方和准则来评价聚类的优劣. 为了得到最佳分类，本文选择基于最邻近规则的试探法这一具体算法.

1. 首先我们假设：

(1) 5 个元音音素/a/o/e/i/u/的标准口形图像(已经过预处理)，即 5 个样本的特征向量依次为 $S_1, S_2, S_3,$

S_4,S_5;

(2) 输入一任意口形图像,经预处理后,得到待识别的新图像,其特征向量为 X;

(3) 选取非负阈值 T 为 5 个样本特征向量之间平均距离的三分之一.

2. 然后,按照下面的公式分别计算 X 与 S_1,S_2,S_3,S_4,S_5 之间的距离 $D(X,S_j)$,$j=1,2,3,4,5$.

$$D(X,S)=\sum_{i=1}^{n}\omega_i x_i s_i \Big/ \sqrt{\Big(\sum x_i^2\Big)\cdot\Big(\sum s_i^2\Big)}.$$

其中(1) X 为待识别的特征向量,S 为样本的特征向量;

(2) ω_i 为加权系数,它的取值由各种特征在识别过程中所起的作用而定,对于较重要的特征,其加权系数就较大;反之,其加权系数就较小.但必须满足 $\sum_{i=1}^{n}\omega_i=1$;

(3) 以上各式中,n 都表示特征个数,或特征分量的个数.

3. 计算完毕,进行比较.若 $|D(X,S)|<T$,且 $D(X,S_k)=\min D(X,S_j)$,其中 $j=1,2,3,4,5$,$k\in\{1,2,3,4,5\}$,则判定 X 属于 S_k 类,即输入音素是 S_k 所代表的元音音素;否则,判定输入音素不是元音音素.

在上述过程中,如遇到边缘特征不明显、较模糊和难于识别的口形图像,则运用歧义图形理解技术进行处理,详见文献[13].

4　算法实现

下面详细叙述本算法的具体实现,并对实验结果加以分析.

1. 存储标准口形并接收口形图像:固定摄像机与说话人之间的距离,在一般条件下(既未采取隔音措施,也无特殊光照射),用摄像机从左右两侧同时对准说话人录像,获取 5 个元音音素的标准口形图像,共 10 幅,并采用.tiff 文件格式和读写模块.将标准口形依次标注为 a_l,a_r,o_l,o_r,e_l,e_r,i_l,i_r,u_l,u_r.然后,对实时说话人,再次通过摄像机摄取其单个口形图像,也存入.tiff 文件中.

2. 图像预处理:现在我们已得到了口形的信息,为了更好地进行特征提取,首先通过低通空间滤波器减少图像中的随机噪音,再利用拉普拉斯边缘增强的方法增强边缘,进行边缘检测预处理.

3. 特征提取:根据前面所讲述的特征选择和提取的基本思想,现在开始进行特征提取.

(1) 从新图像坐标原点开始,行扫描,记录每行像素灰度值最大两点位置 (i,j_1),(i,j_2),计算 $|j_1-j_2|$,则 $\max(|j_1-j_2|)$ 即为展唇度,记作 β;

(2) 从新图像坐标原点开始,列扫描,记录每列像素灰度值最大两点位置 (i_1,j),(i_2,j) 计算 $|i_1-i_2|$,则 $\max(|i_1-i_2|)$ 就是开口度,记作 γ;

(3) 计算新图像边缘内像素点的个数,将其作为口形面积 S,则

$$S=\sum_{i=m}^{n}(|j_{i_1}|-|j_{i_2}|),\ \text{且}\ j_{m_1}-j_{m_2}=j_{n_1}-j_{n_2}=0,\ n>m.$$

采用双目检测技术中由 Levine[14] 提出的模板窗口方法,对新图像进行空间检测,分别提取 5 个元音的圆唇度,记作 α.

具体方法为:设 F_l 和 F_r 分别表示某一元音的左右两幅灰度图像,$f_l(i,j)$ 和 $f_r(i,j)$ 为其对应的灰度函数,则以 (i,j) 为中心,取 $(2u+1)\times(2v+1)$ 窗口作为对应点匹配的基本数据,并在其上作如下相关测度计算:

$$\rho(\Delta d)=\frac{1}{(2u-1)(2v+1)}\sum_{\zeta=i-u}^{i+u}\sum_{\eta=j-v}^{j+v}\frac{f_l(\zeta,\eta)f_r(\zeta,\eta+\Delta d)-\mu_l(i,j)\mu_r(i,j+\Delta d)}{\delta_r(i,j+\Delta d)\delta_l(i,j)}$$

其中 Δd 为估计的位移视差,主要是沿着观察点联线方向取值;μ_l,μ_r,δ_l,δ_r 分别为

$$\mu_l(i,j)=\frac{1}{(2u+1)(2v+1)}\sum_{\zeta=i-u}^{i+u}\sum_{\eta=j-v}^{j+v}f_l(\zeta,\eta),$$

$$\mu_r(i,j+\Delta d)=\frac{1}{(2u+1)(2v+1)}\sum_{\zeta=i-u}^{i+u}\sum_{\eta=j+\Delta d-v}^{j+\Delta d+v}f_r(\zeta,\eta),$$

$$\delta_l(i,j)=\frac{1}{(2u+1)(2v+1)}\sum_{\zeta=i-u}^{i+u}\sum_{\eta=j-v}^{j+v}(f_l^2(\zeta,\eta)-\mu_l^2(i,j)),$$

$$\delta_t(i,j+\Delta d)=\frac{1}{(2u+1)(2v+1)}\sum_{\zeta=i-u}^{i+u}\sum_{\eta=j+\Delta d-v}^{j+\Delta d+v}(f_t^2(\zeta,\eta)-\mu_t^2(i,j)).$$

有了 $\rho(\Delta d)$的计算，然后取 Δd^*，使得

$$\rho(\Delta d^*)=\max_{\Delta d}\{|\rho(\Delta d)|\},$$

则 Δd^*就是该元音的圆唇度.

4. 根据统计模式识别中基于试探的聚类算法，对输入口形进行聚类分析；由分析结果判定是否元音，是何元音. 算法框图如图1所示.

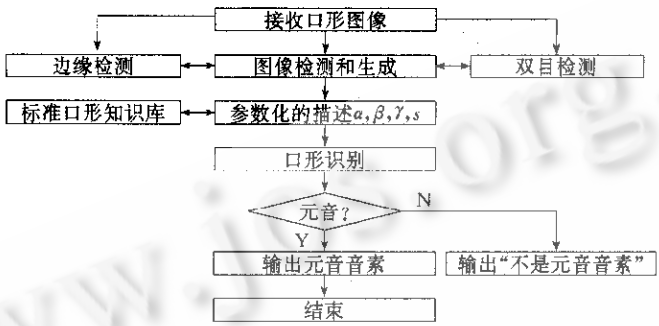

图1 算法示意图

5 实验结果分析

识别程序采用C语言，在PC586机器上运行该程序. 从实验结果我们不难看出，用拉普拉斯边缘增强法可以很好地进行边缘增强，双目检测技术较成熟，而基于最邻近规则的试探法计算简单，可以较快地获得合理的聚类结果. 用此算法进行单个口形的元音音素识别，成功率可达80%. 具体情况见表2、3.

表2 运行程序所得标准口形知识库

特征名 \ 音素名	a	o	e	i	u
圆唇度	1	4	1	1	2
展唇度(左)	72	27	81	80	30
展唇度(右)	76	29	87	85	35
开口度(左)	42	13	20	10	7
开口度(右)	45	15	23	12	9
口形面积(左)	2 325	260	1 225	650	180
口形面积(右)	2 480	296	1 487	776	236

表3 实验数据分析统计表

测试序号	测试口形数	每一口形所取特征数	识别出口形数	识别率(%)
1	20	7	17	85
2	20	7	16	80
3	20	7	15	75
4	20	7	17	85

6 结 论

本课题在语音学的研究基础上，按照图像分析技术、模式识别技术和歧义图形理解技术，对单个口形图像进行形状识别，特别是对5个元音音素的口形进行了识别，实验正确率达80%以上. 这无疑为语音识别提供了一种有用的辅助方法和技术. 当然，如何识别口形序列、如何在音节水平上确认音节语音则是有待于进一步深入研究的课题.

参考文献

1 Lewis J P, Parke F I. Automated lip-synch and speech synthesis for character animation. In: Proceedings of Human Factors Computer System Graphics Interface'87. Toronto, Canada, 1987. 143～147

2 Wu Jian-tong, Tamura S, Mitsumoto H *et al*. Neural network vowel-recognition jointly using voice features and mouth shape image. Pattern Recognition, 1991,24(10):921～927

3 Hight R L. Lip-reader trainer. Tohns-Hopkins Apl. Technical Digest, 1982,10(3):213～237

4 Petajan E D. Automatic lip-reading to enhance speech recognition. Procedure IEEE Computer Association Conference Computer Vision Pattern Recognition, 1985,12(3):44～47

5 Matsuoka K, Furuya T, Kurosu K. Speech recognition by image processing of lip movements. Journal of Association Instrument Control Engineers, 1986,22(10):67～74

6 Uchimura K, Michida J, Tokou M *et al*. Discrimination of Japanese vowels by image analysis. Transactions of Institute of Electronics, Information Communication, Engineers J 71－D, 1988,12(12):2700～2702

7 Kurosu K, Furuya T, Matsuoka K *et al*. Word-recognition by mouth shape and voice. In: Proceedings of the 1st Symptom Advanced Man-Uech. Interface Through Spoken Language. Tokyo, Japan, 1988. 205～206

8 徐世荣.普通话语音发音示意图解.上海:上海教育出版社,1979
(Xu Shi-rong. An Illustration of Chinese Speech Pronunciation Schematic Diagram. Shanghai: Shanghai Education Press, 1979)

9 邓斯 P B,平森 E N.言语链——说和听的科学.北京:中国社会科学出版社,1983
(Dense P B, Pingson E N. The Language Link——Speaking and Listening's Science. Beijing: Chinese Social Science Press, 1983)

10 林焘,王理嘉.语音学教程.北京:北京大学出版社,1992
(Lin Tao, Wang Li-jia. Phonetics Lectures. Beijing: Beijing University Press, 1992)

11 徐建华.图像处理与分析.北京:科学出版社,1992
(Xu Jian-hua. Image Processing and Analyzing. Beijing: Science Press, 1992)

12 沈清,汤霖.模式识别导论.长沙:国防科技大学出版社,1992
(Shen Qing, Tang Lin. Introduction of Pattern Recognition. Changsha: National University of Defense Technology Press, 1992)

13 周昌乐,施项君.歧义图形机器理解初步.计算机软件与应用,1996,13(2):36～41
(Zhou Chang-le, Shi Xiang-jun. An introduction of understanding different meanings' graph with machine. Computer Software and Application, 1996,13(2):36～41)

14 Levine M D, O'Handley D A, Yagi G M. Computer determination of depth maps. Computer Graphics and Image Processing, 1973,13(2):134～150

A Method of Mouth-shape's Shape Recognition Forward to Chinese Speech Recognition

ZHONG Xiao[1] ZHOU Chang-le[2] YU Rui-zhao[1]

[1](*Intelligent Software Laboratory Department of Computer Science Zhejiang University Hangzhou* 310027)
[2](*Vision & Audition Laboratory Department of Computer Science Hangzhou University Hangzhou* 310028)

Abstract This paper presents in detail a statistical method and implementation of mouth-shape's shape recognition and gray level based on perfect image features of mouth shapes at the level of vowels recognition. The results of the recognition experiments approximately accord with the reasoning evaluations. The recognition rate is over 80% for five vowels' mouth-shapes. It proposes a beneficial associate approach for language's voice wave recognition.

Key words Chinese speech recognition, mouth-shape features abstracting, lip-reading, statistical pattern recognition.