一种基于声调规范模型的声调变换方法

薛　健[1]　蔡莲红[2]
[1]清华大学继续教育学院,北京 100084)
[2]清华大学计算机科学技术系,北京 100084)

摘　要　该文利用固定点频率分析提取基音频率(F0),建立归一化线性多项式声调模型。参考男声、女声基音频率的分布和五度标调法,提出了一套汉语声调的规范模型,在此规范模型的基础上,实现了汉语语音声调变换。实际测听表明,经此模型变换的声音达到预期效果。

关键词　基音频率(F0)　声调规范模型　声调变换

文章编号 1002-8331-(2005)10-0040-04　文献标识码 A　中图分类号 TP391.42

A Tone Transformation Method Based on Standard Tone Model

Xue Jian[1]　Cai Lianhong[2]
[1](School of Continuous Education,Tsinghua University,Beijing 100084)
[2](Dept.of Computer Science and Technology,Tsinghua University,Beijing 100084)

Abstract: In this article,method of Fixed Point Frequency Analysis was used to extract Fundamental Frequency (F0)as well as establish the normalized Liner Polynomial Tone Model.A set of standard tone models was proposed in light of the distribution of male voice,female voice fundamental frequency and five degrees of tone marked,by which way we have realized tone transformation of Chinese speech.By detecting of the hearing test,anticipated performance has been reached.

Keywords: fundamental frequency (F0),standard tone model,tone transformation

1　前言

随着语音技术的不断发展,大量研究成果已经应用于现代社会的各个领域,特别是近几年 TTS (Text To Speech)系统通过不断增加语音数据库容量，提高系统合成输出语音的质量。不断扩大的语音库容量已成为 TTS 系统应用中的障碍，为了减少语音库数据量和改变合成语音的自然度,声调变换作为一种有效的方法,越来越受到人们的重视。另一方面,目前大多数 TTS 系统受语音库容量和语料的限制,一般只能合成输出有限的一种或少数几种音色输出,通过声调变换就可以在不增加语音库容量的情况下，为合成语音的多音色输出提供一种简单、实用的方法。

汉语是一种声调语言，汉语声调模型一直受到人们的重视,在文[5]中杨顺安提出了一种归一化线性多项式声调模型,用一个实系数 4 次多项式近似模拟一声到四声的不同曲线。许毅在文[2]中提出的一种指数化递推声调模型,利用递推方程实现对目标声调的逐步逼近。

文献[5]提出的线性多项式声调模型,经过实验统计和计算,提出了一套简单、实用的汉语声调规范模型,为实现汉语声调变换提供了一个确定的框架。实现汉语声调变换时,首先将输入语音基音(F0)信息与声调规范模型进行最佳匹配,经过对两者之间细节差异的处理,实现由输入语音基音到最佳声调规范模型的变换。再由声调规范模型之间的转换,实现给定汉语语音声调的多种变换。该方法既保持了源语音的绝大部分特征,又实现了语音声调的各种变换。

2　固定点频率分析法提取 F0

基音频率(F0)的提取采用文献[1]中提出的理论,利用固定点频率分析法正确估算语音信号基音频率,具体算法为：

第一步,滤波器设计。利用 Gabor 函数设计一个带通滤波器,用来分离和选取基音频率成份。

第二步,频率固定点计算。利用从滤波器中心频率到滤波器输出瞬时频率的不同特性，对每倍频的声音用 24 个滤波器组进行滤波,提取候选基频成份对应的固定点。

第三步,载波噪音(C/N)比率计算。引入振幅在 $0<\varepsilon<<1$ 之间的正弦噪音成分,计算每个固定点对应的噪音成分相对振幅和噪音能量,进一步计算得到这些固定点对应的 C/N 比率值。

第四步，固定频率点选择。选择 F0 的标准是当最大 C/N 比率等于或大于 20dB 时,选择能提供最大 C/N 比率的基频成分为固定点。

第五步,周期分解和 F0 优化。先对 F0 进行自适应短时傅立叶变换(STFT),再对 F0 信息和 F0 的导数进行抛物线时间对称轴规整，对求出的对应谐波成分的固定点进行固定点分解,这些固定点的瞬时频率是完整的,用它们 C/N 信息提供 F0 估算具有最小估算误差，估算得到的谐波成分的 C/N 比率也提供可用于语音再合成的源信号周期控制信息。详细算法介绍见文献[1]。

基金项目:国家 863 高技术研究发展计划项目(编号:2002AA117010);国家自然科学基金项目(编号:60275014)

作者简介:薛健,男,清华大学继续教育学院计算机应用专业研究生课程进修班学员,主要研究方向为语音变换和语音合成。蔡莲红,女,教授,博导,主要研究方向为多媒体技术和语音合成。

3 汉语声调的规范模型

基音频率（F0）是指浊音的声带振动频率，F0的不同轨迹称为声调[3]。声调的高低和变化范围因人而异，一般男声大致在100—200 Hz，女声大致在150—300 Hz[4]。基于这一统计结果，可以假设规范基音频率（F0）取值范围。如果认为五度标调法中五度的对应值均分基频规范取值域，则男声五度对应的规范基音频率平均值分别为：100、125、150、175和200 Hz，女声五度对应的规范基音频率平均值分别为：145、185、225、265、305Hz。

杨顺安在文[5]中提出的归一化线性多项式声调模型为：

$$F_0(t)=\log^{-1}[f_c+f_d\cdot f_{0i}(t)] \quad (1)$$

其中：t为归一化时长；

$i=\{1,2,3,4\}$分别表示汉语的阴平、阳平、上声和去声4个声调；

f_c为体现声调高低的中值频率；

f_d为反映声调基频变化的调域；

$f_{0i}(t)$为调形函数，4种调形函数分别为：

$$f_{01}(t)=0.453+0.295t-1.456t^2+2.574t^3-1.468t^4$$

$$f_{02}(t)=0.011+0.16t-0.913t^2+3.751t^3-2.56t^4$$

$$f_{03}(t)=-0.155+0.246t-7.845t^2+16.36t^3-8.72t^4$$

$$f_{04}(t)=0.463+1.205t-5.584t^2+6.437t^3-3.387t^4$$

由式（1）可产生出4种声调的基音频率序列，以基音频率序列可转换为基音周期序列。

基于以上规范基音频率（F0）取值范围假设和文[5]中提出的归一化线性多项式声调模型，该文提出的汉语声调规范化模型如下：

由（1）式可得：

$$F_{0i}(t)=f_c+f_d\cdot f_{0i}(t) \quad (2)$$

其中：t、i、f_c、f_d物理意义同（1）式，$f_{0i}(t)$为调形函数，其统一表达式为：

$$f_{0i}(t)=a_i+a_2t-a_3t^2+a_4t^3-a_5t^4 \quad (3)$$

将（3）式代入（2），经整理可得到汉语声调的统一表达式为：

$$F_{0i}(t)=a_i+b_it-c_it^2+d_it^3-e_it^4 \quad (4)$$

根据规范基音频率（F0）取值范围和（4）式，经过实验统计和计算，可得到汉语声调四声规范模型分别为：

一声（阴平）规范模型为：

$$F_{01}(t)=a_1+b_1t-c_1t^2+d_1t^3-e_1t^4 \quad (5)$$

其中各对应调型系数取值如表1，其中男声一声（阴平）规范模型曲线如图1所示。

表1 声调规范模型一声（阴平）系数取值表

	调型	a_1	b_1	c_1	d_1	e_1
男声	11	100.000				
	22	125.000				
	33	150.000	29.500	145.600	257.400	146.800
	44	175.000				
	55	200.000				
女声	11	145.000				
	22	185.000				
	33	225.000	29.500	145.600	257.400	146.800
	44	265.000				
	55	305.000				

二声（阳平）规范模型为：

$$F_{02}(t)=a_2+b_2t-c_2t^2+d_2t^3-e_2t^4 \quad (6)$$

其中各对应调型系数取值如表2，其中男声二声（阳平）规范模型曲线如图2所示。

表2 声调规范模型二声（阳平）系数取值表

	调型	a_2	b_2	c_2	d_2	e_2
男声	13	100.000	24.000	136.950	535.650	384.000
	14	100.000	36.000	205.425	803.475	576.000
	15	100.000	47.200	269.335	1053.445	755.200
	24	125.000	24.000	136.950	535.650	384.000
	25	125.000	35.200	200.860	785.620	563.200
	35	150.000	23.200	132.385	517.795	371.200
女声	13	145.000	38.400	219.120	857.040	614.400
	14	145.000	57.600	328.680	1285.560	921.600
	15	145.000	76.800	438.240	1714.080	1228.800
	24	185.000	38.400	219.120	857.040	614.400
	25	185.000	57.600	328.680	1285.560	921.600
	35	225.000	38.400	219.120	857.040	614.400

三声（上升）规范模型为：

$$F_{03}(t)=a_3+b_3t-c_3t^2+d_3t^3-e_3t^4 \quad (7)$$

其中各对应调型系数取值如表3，其中男声三声（上升）规范模型曲线如图3所示。

表3 声调规范模型三声（阳平）系数取值表

	调型	a_3	b_3	c_3	d_3	e_3
男声	313	150.000		910.020	1757.064	
	314	150.000		929.633	1801.236	
	315	150.000		957.090	1855.224	
	413	175.000		1066.920	1889.580	
	414	175.000		1097.516	1945.204	
	415	175.000		1135.172	2009.008	
	513	200.000	24.600	1213.622	2013.916	872.000
	514	200.000		1242.648	2066.268	
	515	200.000		1278.735	2126.800	
	424	175.000		911.589	1758.700	
	425	175.000		935.909	1809.416	
	524	200.000		1074.765	1897.760	
	525	200.000		1103.792	1951.748	
	535	200.000		682.515	1510.028	
女声	313	225.000		1137.525	1986.104	
	314	225.000		1184.595	2072.812	
	315	225.000		1231.665	2159.520	
	413	265.000		1349.340	2157.884	
	414	265.000		1412.100	2257.680	
	415	265.000		1474.860	2362.384	
	513	305.000	24.600	1557.233	2326.392	872.000
	514	305.000		1643.528	2452.364	
	515	305.000		1710.210	2560.340	
	424	265.000		1141.448	1987.740	
	425	265.000		1192.440	2080.992	
	524	305.000		1368.953	2175.880	
	525	305.000		1427.790	2275.676	
	535	305.000		1139.094	1986.104	

四声（去声）规范模型为：

$$F_{04}(t)=a_4+b_4t-c_4t^2+d_4t^3-e_4t^4 \quad (8)$$

其中各对应调型系数取值如表4，其中男声四声（去声）规范模型曲线如图4所示。

由各图表可以看出规范模型比较简单，但它可以明确区分不同声调之间的差异，对不同声调建立了明确的函数表达式，

利用这些函数,为下面的声调变换提供了一种有力的工具。

表 4　声调规范模型四声(去平)系数取值表

	调型	a_4	b_4	c_4	d_4	e_4
男声	31	145.000	120.500	558.400	680.000	338.700
	41	170.000			705.000	
	51	195.000			730.000	
	42	170.000			730.000	
	52	195.000			705.000	
	53	195.000			730.000	
女声	13	220.000	120.500	58.400	619.000	38.700
	14	260.000			659.000	
	15	300.000			699.000	
	24	260.000			699.000	
	25	300.000			659.000	
	35	300.000			699.000	

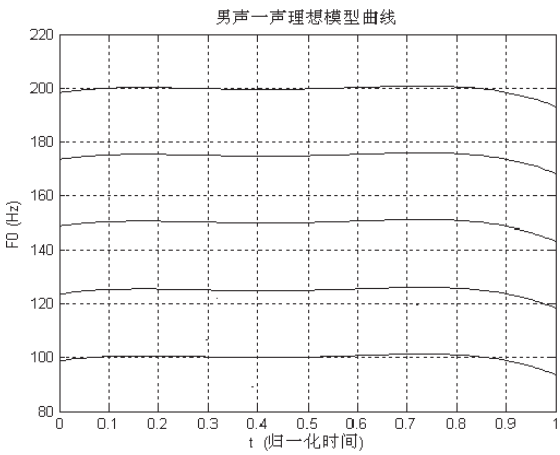

图 1　男声一声(阴平)规范模型曲线

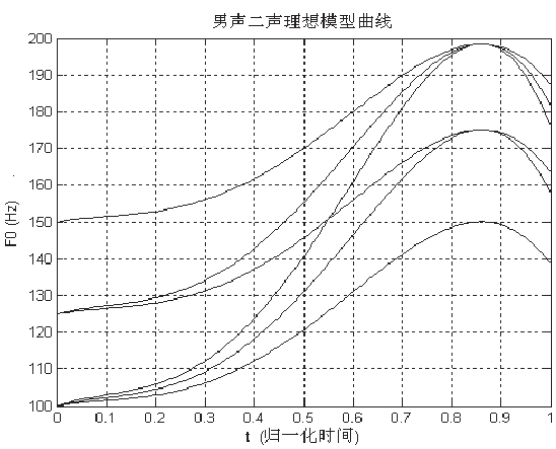

图 2　男声二声(阳平)规范模型曲线

4　声调变换

如果需要对源语音信号进行声调变换,首先要对源语音信号进行基音频率(F0)提取和声调类型识别,根据判别结果与对应规范模型进行最佳匹配,完成匹配后进行源语音到最佳规范模型之间的转换,然后在规范模型之间进行转换,实现声调的各种变换。具体步骤如下:

首先,利用第一节介绍的固定点频率分析法,利用 Gabor 函数设计的带通滤波器,提取基音频率对应的固定点,利用载波噪音(C/N)比率值选择基音频率对应值,利用抛物线时间对

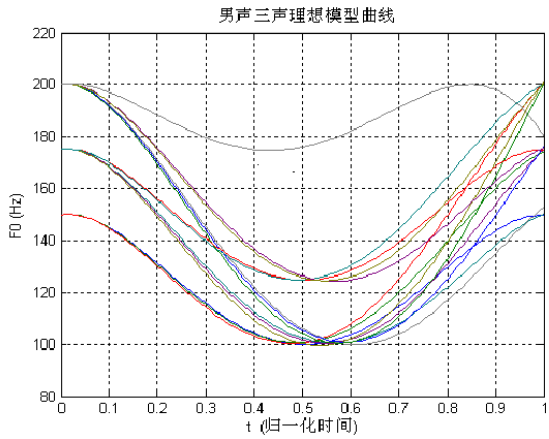

图 3　男声三声(上声)规范模型曲线

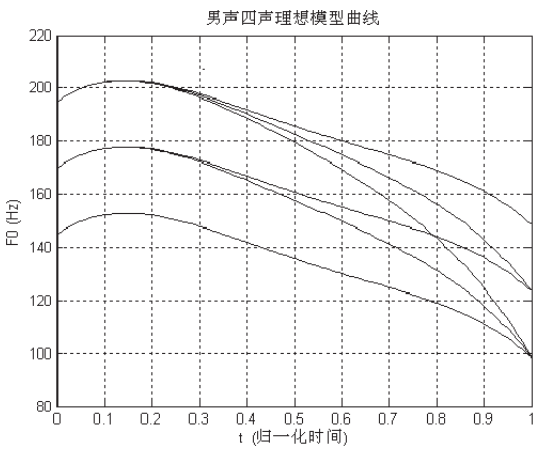

图 4　男声四声(去声)规范模型曲线

称轴规整对基音频率值进行精确化,得到源语音信号对应的基音频率值;

第二步,对得到的基音信息进行声调类型判别。利用上一步得到的基音信息,提取基音频率的平均值、最大值、最小值,提取最大值、最小值对应的归一化时间序列号,采用自适应技术求取最大值和最小值前后一个动态小范围的最大平均值和最小平均值。利用这一技术可以克服由于个别野点造成的局部极值。利用最大平均值、最小平均值、平均值、最大值时间序列值和最小值时间序列值进行组合判断,得到给定语音声调类型。由于该文采用的声调判别方法比较简单,声调类型判别的准确率不是很高,但由于该文只希望得到一个声调曲线的近似趋势,减少最佳匹配时的盲目性,节省运算时间,提高程序效率,其判定结果对声调变换质量没有影响,因此采用这样的简单组合判断可以满足该文声调变换方法的需要。

第三步,根据第二步得到的给定语音声调类型,计算输入语音声调信息与同类声调所有声调规范模型之间的欧氏距离偏差 $D_E F_{0i}$:

$$D_E F_{0i}=\sum_{t=0}^{1}\left[\left(F_{0i}(t)-F_{0i\_ideal}(t)\right)^2\right]^{\frac{1}{2}} \qquad (4)$$

其中 $F_{0i}(t)$ 为源语音基音频率值,$F_{0i\_ideal}(t)$ 为声调规范模型基音频率值。

这里的偏差可以根据需要选取不同量,该文采用欧氏距离,具有较好的判定效果,根据偏差最小原则,判断出最佳匹配规范模型。

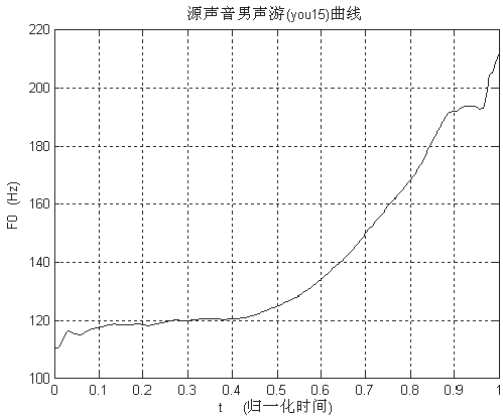

图 5 源语音游 (ou15)基音曲线

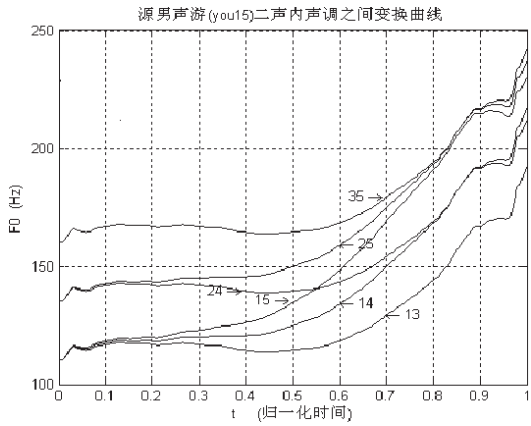

图 6 二声内声调之间的变换基音曲线

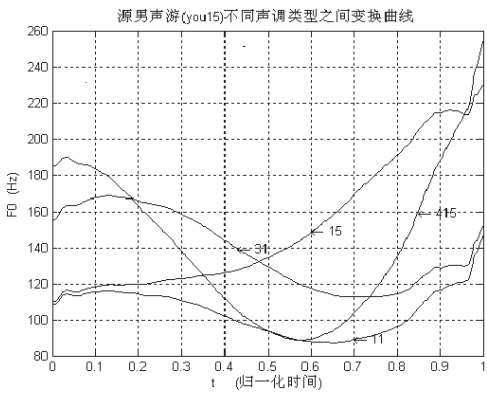

图 7 以 you15 为基准进行不同声调变换曲线

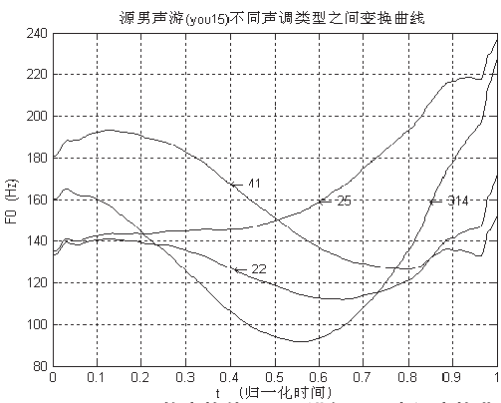

图 8 以 you15 的变换体 you25 进行不同声调变换曲线

第四步,计算源语音基音信号与最佳匹配规范模型对应点参数的细节偏差,从而得到了一个基音信号偏差量 Δf_{0i} (),其计算公式为:

$$\Delta f_{0i}() = f_{0i}() - f_{0i\_ideal}() \quad (0)$$

由 Δf_{0i} () 的集合构成了一个源语音基音信号与最佳匹配规范模型对应点参数的细节偏差向量 ΔF_{0i} (),对其规范模型进行细节偏差处理:

$$F_{oi\_new}() = F_{oi\_ideal}() + \Delta F_{0i}() \quad (1)$$

这样就得到了源语音声调对应的规范声调模型,实现了源语音声调到最佳匹配规范声调模型的变换。经过语音再合成,就可以得到变化声调后的语音,由于最佳匹配几乎包含了源语音的全部声调信息,它们的听觉效果几乎完全相同。

第五步,对判定得到的最佳规范声调模型进行规范声调模型之间的对应映射变换。由于归一化的规范模型之间具有一一对应的映射关系,故可以用其它规范模型代换最佳匹配规范声调模型,即可实现规范模型之间的变换,即有:

$$F_{oi\_ided}() \leftrightarrow F_{0j\_ideal}() \quad (2)$$

对每个新的规范模型之间的变换进行细节偏差处理如下:

$$F_{oj\_new}() = F_{oj\_ideal}() + \Delta F_{0i}() \quad (3)$$

这样就得到了一个基于源语音基音频率变换产生的新基音频率信号,经过语音再合成,就可以得到各种变化声调后的语音。

5 实验结果

读者采用男声游 (ou15)为源语音 如图 5)信号进行了各种声调变换,由图中可以看出其基音具有明显的二声特征,由最大值 (12Hz)、最小值 (12Hz)和幅度差 (00Hz)可以得到游 (ou15)的特征。

图 6 表示了源语音游 (ou15) 声调在二声之间的变化 (ou15→you13、you14、you24、 you25、 you35),由图可以看出变换后的游 (ou15)声调曲线与源语音声调曲线十分相似,其听觉效果也几乎完全相同。其它变换曲线具有明显频率高低差异,经过听觉测试同源语音有明显的高低区别。

图 7 显示了源声音游 (ou15) 声调变换产生的最佳匹配规范声调模型 (ou15), 以及经过不同变换产生 4 种声调 (ou11→you15→you415→you31)的变化。由图可以看出变换产生的一声、三声和四声基音曲线,具有明显的不同声调特征,通过听觉测试有明显的四声变化区别。

图 8 显示了源声音 (ou15)变换产生的最佳匹配规范模型 (ou15), 再由规范基音模型产生不同四声变换 (ou22→you25→you314→you41)。由图可以看出变换产生的一声、二声、三声和四声基音曲线,具有明显的不同声调特征,通过听觉测试有明显的四声变化区别。

6 结论

该文提出的汉语声调变换方法, 基于 straight 语音系统为平台,对各种输入语音进行了多种声调变换,由各输出语音基音频率曲线图,实际人工听音测试和使用清华大学计算机系人机交互实验室语音标注软件 VisualSpeech 对合成变调语音进行验证,各项测试结果表明该文提出的基于基音频率规范模型

(下转 85 页)

其中 m、m_i、m_{ijk} 和 m_j 分别表示所有训练样本数、属于 C_i 类的训练样本数、属于 C_i 类取值 v_{ik} 的训练样本数和第 j 个特征的量化等级。

3 实验结果

由于目前国内尚无通用的用于训练和测试切分结果的数据库，该文采用超星数字图书馆和国家图书馆扫描的书籍图像作为训练和测试图像。首先，训练图 2 中的 3 个用于提取特征的分类器；然后，从训练图像中挑选 4 类文字（仅字类、英文数字类、标点类、部件类）的训练样本；之后，用训练样本的特征分布估计部件条件概率特征 $P(C_4|X_{\text{形状结构特征}})$ 和贝叶斯分类器的条件概率分布函数，最终完成分类器设计。连写的英文或数字不易切开，而且也没有必要；因为只要判断出它是连写的英文或数字，就可以在后续处理中简单地把它们切开。所以在挑选英文数字类训练样本时，也挑选了一部分连写的英文或数字作为训练样本。在估计 $\hat{P}(C_4|X_{\text{形状结构特征}})$ 时，由于公式（6）中的 $P(X_{\text{形状结构特征}})$ 不易准确估计，而 $P(C_4)$ 是常数，所以实际应用中用 $\hat{P}(X_{\text{形状结构特征}})|C_4)$ 替代 $\hat{P}(C_4|X_{\text{形状结构特征}})$。

用超星数字图书馆和国家图书馆的测试图像测试本文的方法。图 3 是英文、数字相对较少的图像，对它的切分和文字类别判断完全正确。对于英文相对较多的图 4，同样取得了较好的结果。但是对于连写的“93”，贝叶斯分类器错把它们当成汉字进行切分的，而不是按连写的数字把它们切分在一起的。虽然对于切分在“9”和“3”之间的候选位置，英文数字和部件分类器给出了正确的判断，但受其它特征的影响，最终的切分位置还是选在了“3”之后。这说明简单地把所有特征组合在一起的方法是有局限性的，它们需要按一定的结构和层次关系结合在一起。这也是目前正在研究的问题。

图 3 测试图像 1 的切分结果

4 总结

论文针对实际的混排文档图像，提出一种基于贝叶斯分类器的统计学习方法切分文字，并实现文字类别判断。虽然笔者只是简单地把所有特征组合在一起，然后用贝叶斯分类器选择最佳切分位置和判断文字类别，但却取得了较好的效果。如何把特征按一定的结构和层次关系结合在一起，这是目前正在研究的问题。解决这个问题将会进一步提高切分和文字类别判断

（a）切分结果

（b）判断为汉字类的文字

（c）判断为英文数字类的文字

（d）判断为标点类的文字

图 4 测试图像 2

的准确性。

基于统计学习理论的方法具有结构简单、计算量少、易于扩展功能的特点，在切分和文字类别判断应用领域应该是非常有发展前景的。（收稿日期：2005 年 1 月）

参考文献

1.Casey Richard G，Lecolinet Eric.Survey of methods and strategies in character segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，1996，18（7）：690~706

2.R L Hoffman，J W McCullough.Segmentation Methods for Recognition of Machine-Printed Characters[J].IBM J Research and Development，1971，153~65

3.G Nagy.Twenty Years of Document Image Analysis in PAMI[J].IEEE Transactions on PAMI，2000，38~62

4.Kohavi R，Becker B，Sommerfield D.Improving Simple Bayes[R].（technical report）Data Mining and Visualization Group，Silicon Graphics Inc，Mountain View，CA.ftp：//starry.stanford.edu/pub/ronnyk/impSBC.ps.Z

5.J Guo，N Sun et al.Algorithm for recognition of handwritten characters using pattern transformation with cosine function[J].IEICE Trans，1993，J76-D-II（4）：835~842

6.马少平，夏莹，朱小燕等.汉字识别系统的误识模型[J].清华大学学报，1999，38：108~111

7.吕岳，施鹏飞，张克华.基于汉字结构特征的自由格式手写体汉字切分[J].电子学报，2000，28（5）：102~104

（上接 43 页）

进行的语音声调变换方法，实现方法简便实用，变换结果声调调型特征明显，可以认为是一种汉语语音声调变换的有效实用方法。绝大部分变换结果具有较好的声音质量，个别变换语音有失真，需要进一步改进算法予以改善。

致谢

笔者向能及时提供讨论的清华大学计算机系人机交互实验室的黄德智博士、杨鸿武博士和其它同事表示感谢。

（收稿日期：2004 年 12 月）

参考文献

1.Hideki Kawahara，Haruhiro Katayose，Alain de Cheveign et al.Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity[C].In：Proceedings of EUROSPEECH'99，Vol6，1999：2781~2784

2.Ching X Xu，Yi Xu，Li-Shi Luo.A pitch target approximation model for F0 contours in mandarin[C].In：Proceedings of Icphs99，san francisco，1999：2359~2362

3.杨行峻，迟惠生等.语音信号数字处理[M].北京：电子工业出版社，1995：20

4.林焘，王理嘉.语音学教程[M].北京：北京大学出版社，1992：125

5.吴宗济，林茂灿等.实验语音学概要[M].北京：高等教育出版社，1989：328