

# 基于听觉模型的汉语耳语音声调检测

陈雪勤, 赵鹤鸣

(苏州大学电子信息学院, 江苏苏州 215021)

**摘要:** 从听觉感知出发, 分析了听觉外周模型对于语音激励的主要响应过程, 采取听神经平均发放率为声调感知线索, 提出了一种汉语耳语音声调的识别方法. 其理论基础是听神经发放信息是听觉中枢的唯一信息来源, 它是对于语音激励中声强、频谱、共振峰等多种特征的综合反应, 因此适合用作耳语音的声调特征. 采用BP神经网络对大量汉语元音耳语音四声样本进行训练、识别, 得到65.1%的平均识别率, 达到了改善汉语耳语音声调识别效果的目的.

**关键词:** 声调检测; 汉语耳语音; 听觉模型; 听神经平均发放率

**中图分类号:** TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2009) 04-0864-04

## Perceiving of Tone in Whispered Chinese Based on Auditory Model

CHEN Xue-qin, ZHAO He-ming

(School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu 215021, China)

**Abstract:** Based on the analysis of the response of a peripheral auditory model for speech stimulation, the average firing rate of auditory nerves is chosen as the cue for whispered tone. Thus a method for whispered Chinese tone perceiving is proposed. The underlying principle is based on the fact that auditory nerve is the only source of information for central auditory system and it responds to several types of acoustic stimulus such as intensity, formant, etc. Therefore the average firing rate of auditory nerves is a suitable characteristic for the tone of whispered speech. The BP artificial neural network was trained by these proposed parameters to achieve tone recognition. Experiments are performed on a lot of Chinese whispered speech data and the average correct rate reaches 65.1%, which shows that the proposed method is effective for improving the performance of whispered Chinese tone perceiving.

**Key words:** tone detection; whispered Chinese; auditory model; the average firing rate of auditory nerves

## 1 引言

耳语音调感知的研究对于耳语音的处理如增强、识别等具有重要意义<sup>[1,2]</sup>. 当正常语音的声调主线索—基频的研究向高精度方向发展时<sup>[3,4]</sup>, 耳语音由于传统声调的感知线索—基频的缺失, 关于耳语音是否有声调, 以及如何感知其声调依然成为研究热点. 长期以来持耳语音声调信息可感知观点的学者们采用主观视听和客观测试的手段验证耳语音声调可感知和探索耳语音声调信息的载体, 得到了一些结论. W. Meyer-Eppler<sup>[5]</sup>利用可视语音分析仪经实验证明正常语中音调随基频变化的特征在耳语音中由某些共振峰的偏移所替代. Martin Kloster Jensen<sup>[6]</sup>更倾向于直接的感知测试. 文献[6]对四种有调语言—挪威语、瑞典语、斯洛文尼亚语和汉语普通话进行主观测听实验, 得到的识别正确率从53%到100%不等. 文献[7]通过对所有汉语耳语音进行声调主观测试, 得出人耳对于耳语音声调的平均分辨率为62.1%的结论. 文献[5~7]的共同点是通过主观实验证明了耳语音声调是可感知的. 近年来, 耳语音声调的声学特征研究又有所进展, Gao-man<sup>[8]</sup>指出音长及幅度是汉语普通话声调的重要声学特征, 对于汉语孤立字, 按照感知难度

由小到大排序, 分别为第三声、第四声、第一声、第二声. 文献[9]指出幅值包络、音长、声道面积、声道长度、及共振峰都可以是汉语耳语音的声调特征. 并进一步结合人耳的听觉特征, 采用响度加权的32个Mel频段的对数幅值包络加音长的特征量, 声调平均正识率可达到64.9%. 已有方法主要着眼于语音的声学特征的提取, 总体识别率并不理想. 这促使人们进一步去研究人类听觉模型在耳语音声调感知中的作用.

本文着眼于听觉感知机制, 用听神经平均发放率来研究汉语耳语音的声调感知, 并以此为基础, 提取合适的耳语音声调特征参数, 利用神经网络技术得到六个汉语耳语音元音不同声调各自的识别率数据.

## 2 听神经发放率与声调

听神经纤维将耳蜗内毛细胞与听觉中枢神经系统联接起来, 它是听觉中枢的唯一信息来源. 每条听觉神经纤维与基底膜的一个特定部位相对应, 并在一个特定频率上发放. 神经的激发频率与有多少神经元参与有关, 声强愈高, 神经元愈多, 而神经元的激发频率亦愈快. 正常状态下, 放电率与声刺激强度关系呈形. 同时已有的研究表明<sup>[10]</sup>, 听神经纤维具有与刺激同步发

收稿日期: 2007-10-23; 修回日期: 2008-10-23

基金项目: 国家自然科学基金(No. 60572076); 江苏省高校自然科学基金(No. 05KJB510113)

放的能力,听觉神经纤维能够对共振峰的刺激谐波锁相或同步.因此,听神经发放与声刺激的强度、频谱、共振峰等信息密切相关.

由于耳语音缺少基频这一最重要的声调感知线索.因此其声调的感知源于其他诸如语音信号强度、频谱、共振峰等本处于辅助作用的线索中.而听觉神经发放信息是对于声刺激多种信息的综合反应,将其作为声音信号的声调感知线索具有较强的优势.

### 3 听觉外周模型及特征提取

听觉外周计算模型包括对外耳、中耳、耳蜗基底膜、内毛细胞及听觉神经纤维的模拟,该部分可获得语音信号的信息表示,然后将信息传送给中枢神经<sup>[10,11]</sup>.

声音经外耳、中耳放大及阻抗变换后传到内耳,变成了沿耳蜗基底膜传播的压力波,并引起基底膜的上下运动和耳蜗覆膜的剪切运动.基底膜对于声音具有频率选择性,反映到人耳即为听到的声音高低与声音的频率呈非线性关系.可取符合听觉特性的 Gammatone<sup>[12]</sup>滤波器组进行模拟,每个滤波器可表示为:

$$G_m(t) = a \cdot t^{n-1} e^{-b_m t} \cos(2 f_m t + \phi_m) \quad (1)$$

其中  $n=4$ ,为滤波器的阶数, $b_m$ 是中心频率, $f_m$ 为在等效矩形带宽(ERB)域上的变换频率,它在该域是等间距分布的.根据语音信号的频率分布特征,令  $f_m$  的取值范围为 20Hz~4kHz,它与  $b_m$  的关系为:

$$b_m = 1.019 \text{ ERB}(f_m) \quad (2)$$

基底膜的振动刺激了毛细胞,再由毛细胞将声波振动的机械能转换成电能.其间内毛细胞检测耳蜗覆膜的剪切运动,并引起本元神经元的发放.内毛细胞以及与听觉神经相连的突触区域将作半波整流、非线性饱和和抑制、短时自适应等反应.将内毛细胞与听觉中枢神经系统联接起来的是听觉神经纤维,并且它是听觉中枢的唯一信息来源.本文采用 Meddis 提出的内耳毛细胞函数模型<sup>[13]</sup>.

$$k(t) = \begin{cases} gdt[x(t) + A]/[x(t) + A + B], & [x(t) + A] > 0 \\ 0, & [x(t) + A] < 0 \end{cases} \quad (3)$$

$$\frac{dq}{dt} = y[1 - q(t)] + \kappa(t) - k(t)q(t) \quad (4)$$

$$\frac{dc}{dt} = k(t)q(t) - lc(t) - \kappa(t) \quad (5)$$

神经发放的概率则可表示为:

$$p(F) = hc(t)dt \quad (6)$$

以上四个方程组成了整个内毛细胞/突触模型,其中  $g, y, r, l, h, A, B$  是常数,与上述对应的递质变化速率有关,其取值可参阅文献<sup>[14,15]</sup>, $dt$ 为采样间隔.其中  $k(t)$ 为渗透膜的渗透率,由输入信号的幅度决定,它相当于是对基底膜似的输出的信号进行半波整流.该模

型假设毛细胞具有制造递质的功能,它内部所含的可自由释放的递质量以  $q(t)$  表示,而且有  $y[1 - q(t)]$  的补充率.突触间隙内包含的递质量以  $c(t)$  表示,它持续的向毛细胞返回的量为  $\kappa(t)$ ,并且还会有  $lc(t)$  的递质量不断丢失.设每一次神经发放事件的发生所带来的神经脉冲发放量为  $Q$ ,否则为  $R$ .则听神经平均发放率可通过对各通道的听神经发放量做短时积分得到:

$$f(m, k) = \frac{1}{N} \sum_{n=1}^N [Q \cdot p + R \cdot (1 - p)] \cdot W(n) \quad (7)$$

其中  $W(n)$  为积分窗函数, $m$  代表第  $m$  帧, $k$  表示第  $k$  个滤波器通道, $N$  表示帧长.

### 4 讨论与分析

为验证上述特征值对于声调感知的有效性,本文对汉语耳语音作测试.语料库在安静的实验室环境录制,采样率 8000Hz,量化位 16 比特,由 10 人(5 男 5 女)产生,每人发/a/ /o/ /e/ /i/ /u/ /ü 六个元音的四声各 10 遍,共 2400 个音.

#### 4.1 特征量的实现及识别效果

分帧后的语音信号  $x(n)$ ,  $n=1 \sim N$ ,经外中耳滤波后经耳蜗基底膜分解为不同频率段信号  $x(1, n) \sim x(K, n)$ ,  $K$  取 64 时可以较好的拟合人耳的频率特性.进一步毛细胞渗透膜的渗透率  $k(t)$ ,毛细胞自由释放的递质量  $q(t)$ ,突触间隙内的递质量  $c(t)$  三变量关系由方程式(3)~(5)表示,神经发放概率  $p$  即可由上述方程组的解  $c(t)$  获得.实验中对模型参数的具体取值如表 1 所示.表 1 中  $f_s$  为语音激励信号的采样率.

表 1 听觉外周计算模型参数表

参数名	$g$	$y$	$r$	$l$	$h$
取值	1660	20	12500	500	10000
参数名	$dt$	$A$	$B$	$Q$	$P$
取值	$1/f_s^{(1)}$	10	15	1	0

为便于仿真,设每一次神经发放事件的发生所带来的神经脉冲发放量为 1,否则为 0.则特征量听神经平均发放率可简化表示为  $f(m, k) = \frac{1}{N} \sum_{n=1}^N p \cdot W(n)$ ,  $m$  代表第  $m$  帧, $k$  表示第  $k$  个滤波器通道.

对每一帧语音利用上述模型提取听神经平均发放率,以听神经平均发放率和(见 4.2 节)作为特征,每个语音的特征量被线性归整为 25 帧,由此构成一个语音的学习样本,输入到 3 层的 BP 神经网络.将语料库中 3 遍 720 个音作为训练样本,其余 7 遍 1680 个音作为测试样本,得到如表 2 所示数据.由表可知女生平均 59.7%,男生平均 62.7%,总平均识别率为 61.2%.

耳语音发音特点决定了耳语音频谱与正常语音相比向高频偏移,而女声的频谱与男声相比向高频偏移.

表 2 六个汉语耳语音元音声调识别率

性别	元音	识别率 (%)				平均 (%)
		声调 1	声调 2	声调 3	声调 4	
男	/a/	75.2	58.3	73.4	98.8	62.7
	/o/	58.6	81.8	92.5	91.6	
	/e/	80.7	66.7	65.2	37.5	
	/i/	60.2	17.5	29.3	81.4	
	/u/	63.2	26.4	42.7	90.1	
	/ü/	56.5	37.4	37.5	84.2	
女	/a/	70.5	55.3	70.2	95.6	59.7
	/o/	54.8	78.7	88.1	90.7	
	/e/	73.6	62.5	62.8	35.5	
	/i/	57.2	15.6	26.6	80.1	
	/u/	60.4	26.3	38.5	87.6	
	/ü/	50.1	33.9	35.4	82.7	
平均 (%)		63.4	46.7	55.2	79.5	

由于 Gammatone 滤波器形状及带宽的分布决定了高频段信息有所损失,以及受到语音样本采样率的限制,本文所用特征量对男性发音者的声调信息更具表现力.表 2 的数据也显示男声的识别率要高于女声.比较表 2 中四声声调识别率,从高到低依次是四声、一声、三声、二声,与文献[8]相比,第三声识别率排序落后.在耳语音四声声调中第三声的平均音长最长,这是在主观听觉中耳语音第三声调识别率排在首位的重要线索,而本文模型中各语音特征量被归整为 25 帧,所用特征量中不带有音长信息,使第三声调识别率排序下降.

## 4.2 各频段的声调信息量分析

正常语音由于具有基音频率及谐波特性,因此在整个频带都携带有声调信息.耳语音则不同,由于基频的缺失,其声调信息的携带者发生变化.实验表明,位于滤波器低频段的听神经平均发放率基本不具备声调信息,而高频段的听神经发放率含有明显的声调信息.

图 1 中 (a)、(b) 显示了耳语音元音/o/ 四声,经听觉外周模型处理,取 64 个 Gammatone 滤波器中低十个(0Hz ~ 538Hz)和高十个(2390Hz ~ 4000Hz)滤波器输出得到的听神经平均发放率.由图中可以看出四声的低通道信息几乎没有区别,即不体现音调信息,而四声高通道信息完全不同,可由该三维图形观察出声调趋势.这也进一步验证了声调信息主要包含在耳语音信号的频谱细节中,而不是在频谱包络中.根据这一理论可对具有声调信息的高通道信息求和,图 1 中 (c) 为各高通道的平均发放率之和,将多维的信息综合至一维,耳语音的声调信息得到了很好的保留.以图 1 (c) 所得的平均发放率和作为特征量,不仅可以大大降低神经网络声调识别运算量,且可更为直观的将耳语音声调信息与听觉模型特征之间的关系表现出来.

## 4.3 分帧对声调识别效果的影响

分帧处理方法是基于语音信号的短时平稳特性,给语音信号处理带来便利的同时,也损失了听觉系统的时间连续性.听觉心理和生理学研究表明,人的听觉系统在对信号处理分析时保留了语音信号流的连续

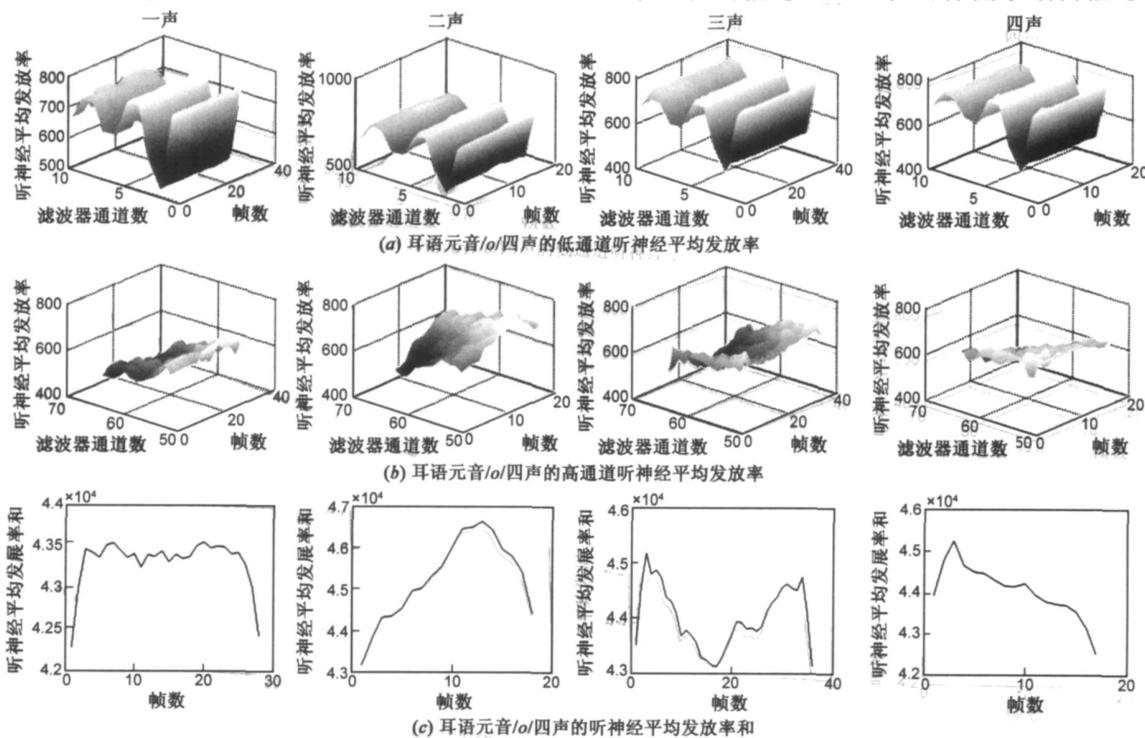


图 1

性,充分利用上下文信息识别、分析语音信号. 本文通过实验分析和讨论了分帧对于耳语音声调识别的影响.

表3 分帧与不分帧情况下的四声声调平均识别率比较

声调	一声	二声	三声	四声
分帧	63.4%	46.7%	55.2%	79.5%
不分帧	66.9%	50.6%	59.2%	83.7%

分帧与否在算法实现过程上有所区别. 分帧处理方法如第 4.1 节所述. 不分帧方法是对一个元音音节,在进入听觉计算模型之前不分帧,外中耳滤波和基底膜带通滤波在时域采用卷积方法实现. 内毛细胞/突触模型针对整个音节进行,最后听神经发放率计算过程中的短时积分长度选择与分帧的帧长一致,以便可比. 表 3 显示了在分帧与不分帧情况下四声声调的平均识别率,不分帧时识别率普遍有所提高,总平均识别率达到 65.1%. 由于保留语音信号的连续性更符合人类听觉处理机制,因而识别效果可取得一定量的改善.

## 5 结论

在耳语发音情况下,它受周围环境的影响甚大,且缺少基频这一重要线索,其声调识别变得尤其困难. 本文着眼于听觉神经机制,集中于听觉外周模型分析,提出基于听觉神经平均发放率的耳语音声调识别方法. 并比较了在分帧与不分帧两种情况下声调识别效果和不同频段的神经发放率中所包含的声调信息. 该方法是对现有的耳语音声调识别特征量在听觉感知分析方面的补充. 实验针对较丰富的耳语音数据,采用神经网络进行训练识别,取得了较为理想的识别效果. 下一步研究工作可着眼于神经发放信息传送给中枢神经以后的处理过程,更加深入的探讨听觉中枢系统对于耳语音声调感知过程,并可结合其他有利于声调识别的声学特征,以期获得更高的耳语音声调识别率.

## 参考文献:

- [1] Morris R W. Enhancement and recognition of whispered speech [D]. USA: Georgia Institute of Technology, 2002.
- [2] Ito T, Takeda K. Analysis and recognition of whispered speech [J]. Speech Communication, 2005, 45(2): 139 - 152.
- [3] 罗亚飞, 鲍长春. 基于 DCT 分带谱熵与信号分解的高精度基音检测算法 [J]. 电子学报, 2007, 35(1): 13 - 22.  
LUO Ya-fei, BAO Chang-chun. Super resolution pitch detection based on band-partitioning spectral entropy and signal decomposition in DCT domain [J]. Acta Electronica Sinica, 2007, 35(1): 13 - 22. (in Chinese)
- [4] 黄海, 潘家强. 基于 Hilbert-Huang 变换的基音周期提取方法 [J]. 声学学报, 2006, 31(1): 35 - 41.  
HUANG Hai, PAN Jia-qiang. Pitch detection method based on Hilbert-Huang Transform for speech signals [J]. Acta Acustica,

2006, 31(1): 35 - 41. (in Chinese)

- [5] Meyer-eppeler W. Realization of prosodic features in whispered speech [J]. Journal of Acoustical Society of America, 1957, 29(1): 104 - 106.
- [6] Martin Kloster Jensen. Recognition of word tones in whispered speech [J]. Word, 1958, 14: 187 - 196.
- [7] 沙丹青, 栗学丽, 徐伯龄. 耳语音声调特征的研究 [J]. 电声技术, 2003, (11): 4 - 7.  
Sha Dan-qing, Li Xue-li, Xu Bo-ling. Study on the characteristics of the tones in whispered Chinese [J]. Audio Engineering, 2003, (11): 4 - 7. (in Chinese)
- [8] Marr-gao. Tones in whispered Chinese: articulatory features and perceptual cues [D]. Thesis of Master, University of Victoria, Canada, 2002.
- [9] Li Xueli, Xu Boling. Tones feature in whispered Chinese [J]. Progress in Natural Science, 2005, 15(3): 285 - 288.
- [10] 吴玺宏, 迟惠生, 王楚. 基于听觉外周模型的语音信号听觉神经表示 [J]. 生物物理学报, 1997, 13(2): 213 - 220.  
Wu Xi-hong, Chi Hui-sheng, Wang Chu. Auditory model based neural representation of speech signal [J]. Acta Biophysica Sinica, 1997, 13(2): 213 - 220. (in Chinese)
- [11] Sachs M B, et al. Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus [J]. Journal of Phonetics, 1988, 16: 37 - 53.
- [12] Patterson R. An efficient auditory filterbank based on the gammatone functions [R]. Annex B of the Svos Final Report: The auditory filter bank, APU Report No. 2341, 1988.
- [13] Meddis R. Simulation of mechanical to neural transduction in the auditory receptor [J]. Journal of the Acoustical Society of America, 1986, 79(3): 702 - 711.
- [14] Meddis R, et al. Implementation details of a computer model of the inner hair-cell/auditory-nerve synapse [J]. Journal of the Acoustical Society of America, 1990, 87(4): 1813 - 1818.
- [15] Meddis R, Hewitt M J. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification [J]. Journal of the Acoustical Society of America, 1991, 89(6): 2866 - 2894.

## 作者简介:



陈雪勤 女, 1974 年 12 月出生于江苏省扬州市. 1997 年毕业于苏州大学, 现为苏州大学博士研究生, 主要研究方向为语音信号处理.  
E-mail: chenxueqin@suda.edu.cn

赵鹤鸣 男, 1957 年 8 月出生于江苏省无锡市. 教授, 博士生导师, 苏州大学电子信息学院院长. 主要研究领域为语音信号处理、神经网络理论与应用.