第 23 卷第 4 期 声 学 学 报 Vol. 23, No.4
1998 年 7 月 ACTA ACUSTICA Jul., 1998

普通话单韵母的听觉规范 *

方 至

(中国科学院心理研究所 北京 100101)

1996 年 12 月 20 日收到

摘要 用普通话若干孤立单韵母作语音材料，有三组性别或年龄不同的发音人，通过声学分析和统计分析，比较了言语识别预加工的三种听觉模型。这些模型由几种不同的听觉变换和规范构成。实验的效果表明，最佳的听觉变换为 Erb，但几种规范之间却没有明显的差别。Chistovich 的频谱重心效应对普通话单韵母也是适用的，韵母的临界距离估算为 3.0 Bark 或 3.5 Erb 。

PACS 数 43.60 43.66 43.70

The auditory normalization of standard Chinese monofinals

FANG Zhi

(*Institute of Psychology The Chinese Academy of Sciences* Bejing 100101)

Received December 20, 1996

Abstract As experimental phonetic materials, this paper uses some standard Chinese monofinals produced isolatedly by three groups of speakers of different sex or age. By means of acoustical and statistical analysis, three auditory preprocessing models of speech recognition, that contain some kinds of both auditory transformations and normalizations, are compared. Experimental results show that Erb is the best one among auditory transformations, but no significant difference is found among these normalizations. Chistovich's SCG effect is verified, the critical distance between standard Chinese finals is estimated as 3.0 Bark or 3.5 Erb.

引言

语音的变异很大，这个特点在语音的声学平面图上看得很明显：一些临近的语音区往往互相重叠，你中有我，我中有你。这给语音识别，尤其是非特定发音人的语音识别带来一些困难。对比人类听者，言语知觉却能巧妙地在语音变异中清楚的感知语音。对语音知觉的这种常性 (constancy)，听觉研究的一些先驱者都曾予以关注。Helmholtz 将它解释为一种 "无意识推论"，Bekesy 在他晚年对听觉研究的展望中着重提出了这个老大难问题。人们有幸看到，当代心理声学和言语声学已定量地对它有了进一层的实验研究。从语音知觉常性现象，可以设想言语知觉中存在着一个变换和规范的过程，能缩小语音变异使语音得以正确知觉。语音变异有两个来源，一类来自不同性别或年龄的发音人声道的差异，一类来自语音语境，语速等的差异。前者限于一个音节内，属于

* 国家自然科学基金资助课题。

内在的 (Intrinsic), 后者是超音节的，属于外在的 (extrinsic)．有关人类听觉系统对语音变异的变换和规范的研究，自 40 年代末 Joos 开始，几经起伏，至今还不能说有定论，但取得的进展是明显的．带有变异的语音，经过听觉系统外周的非线性处理，变换为听觉单位，而后再加以一定线性变换的规范，的确可以收到缩小内在的差异的规范效果．语音的听觉变换利用了心理声学的研究成果．从对数量表，音高的 Mel 量表，到临界频带的 Bark 量表，规范的效果逐步提高．值得注意的是， Moore 和 Glasberg 用改进了实验技术，比较精确地确定了听觉滤波器的带宽．它们接近于 1/6 oct. 比接近 1/3 oct. 的 Bark 量表的带宽为窄，尤其在 500 Hz 以下的低频波段，带宽的变化类似于对数量表，而 BARK 量表这一段的变化是近乎线性的．按此模型确定的听觉单位命名为 ERB, 即等效矩形带宽 (equivalent rectangle band-width) 的缩写． Bark 与 Erb 的上述差异，理应产生不同的变换效果．对听觉变换后语音的进一步规范，如果暂不考虑外在的变异，目前有两种不同的理论模型和方法，它们都有各自的实验根据． Lloyd, Peterson, Miller 等先后都主张一种共振峰比的理论．他们认为，相同的音质来自相同的发音，而且相同的发音产生相同的共振峰比．他们的实验数据也确证，利用元音前三个共振峰之比能减少或接近消除说话人的内在差异． Syrdal 和 Gopal 的观点和方法，利用元音相邻共振峰之差来消除内在的变异，其实验根据是 Chistovich 的频谱重心 (SCG) 和临界距离理论，也同样得到有关实验的支持．

目的是以普通话若干韵母为对象，将上述不同的听觉变换和规范，作为语音识别预加工的几种不同听觉模型进行比较研究，以期对言语规范，言语知觉和言语识别方面的工作，能提供一点有用的资料．

1 方法

1.1 实验语音及其采样

以普通话单韵母 /a/, /e/, /i/, /u/ 作为实验语音．发音人 30 名，包括青年男女及儿童各 10 名，都说普通话，没有听力及言语障碍．发音人每次发音，四个韵母各重复四次， 16 个音随机排列在一张卡片上，按序发出，各音间保持一定间隔，以防连读．采样在一间安静的隔声室内进行．话筒置于发音人嘴前半米处，语音经 B&k2209 声级计接收后，输至带有 SB16 声卡的微机采样．采样频率 11025 Hz ．语音及发音人的上述选择是为了将变异限制在发音人差异引起的范围内，不考虑语境等外在影响所产生的变异．

1.2 语音的声学分析和统计分析

元音或韵母的音质，用它们的前三个共振峰基本上可以表征．对韵母这些特征的分析和提取，借助了 ILS 信号数字处理软件系统．语音波形通过该系统的音高提取分析 (API), 语图显示 (SGM) 等程序，用倒谱和自回归系数 LPC 方法，分别作了音高提取和快速富氏分析，获得了各单韵母的 F_0, F_1, F_2, F_3 等特征值．包含这些数值的 120 (30 发音人 *4 韵母) 个文件，用该系统直方图 (HIS) 程序作统计分析，确定各韵母的均数及其标准差，并以后者衡量语音变异的大小．

1.3 语音特征值的听觉变换及规范

语音 LPC 分析确定的特征值的单位是 Hz ．将比较对它们的三种听觉变换： Mel, Bark 和 Erb ． Mel 的变换，按 Fant 的 TM(Technical Mels) 公式

$$\mathrm{Mel} = (1000/\log 2)\log(f/1000 + 1)$$

Bark 的变换，采用 Traunmuller 对低频波段修改后的公式

$$\mathrm{Bark} = (26.81f/1960.0 + f) - 0.53$$

Erb 的变换，将 Moore, Glasberg 的原有公式，按 Greenwood 提出的 Erb 相当于沿底膜的恒定

距离的假设加以修改

$$\mathrm{Erb} = 18.3\log(0.006046f + 1)$$

以上公式中，f 均为以 Hz 为单位的频率值。

变换为听觉单位后，特征值进一步规范，不论采用共振峰差还是共振峰比，都是相邻两共振峰之间的线性变换，如 $f_1 - f_0$, $f_2 - f_1$, $f_3 - f_2$ 或 f_1/f_0, f_2/f_1, f_3/f_2, 分别替代 f_1, f_2, f_3 。

1.4 模型的评价

的目的在比较几种语音规范的不同听觉模型。怎样评价它们？怎样评价规范的效果？ Disner 曾指出，缩小内在差异引起的变异自然是一个标准，但还不够，还要达到能消除特征图上不同韵母间的重叠现象，保证它们的正确识别。按这样的标准，采用了以下统计分析方法。

(1) 离中系数 (CV): 它以均数为标准差的单位，可以比较不同听觉变换 (单位) 间变异的大小。

(2) 判别分析：可以定量地比较不同模型的韵母识别率和三类发音人的识别率。

(3) 散点分布图：用 ILS 的图形描记 (PLR) 程序，可以在共振峰平面图上以它们的若干个标准差为长短轴的椭圆，直观地显示韵母间有无重叠和发音人变异缩小程度的规范效果。

进行变换，规范及统计分析所用统计软件包括 CSS 和 SPSS 。

2 结果

2.1 韵母变异及听觉变换对它的作用

韵母变异的大小，可以用它的特征值 F_1, F_2, F_3 在 30 名发音人中的离中趋势表征。为了比较经过听觉变换后变异有无缩小及缩小的程度，表 1 列出了各韵母的以 Hz 为单位的声学特征值和以 Bark, Erb, Mel 为单位的听觉特征值的离中系数。

表 1 韵母特征值听觉变换前后的离中系数 (%)

韵母	特征值	Hz	Erb	Bark	Mel
/a/	f_1	11.8	6.4	8.5	8.6
	f_2	6.8	2.9	4.4	4.7
	f_3	11.6	3.8	4.8	6.3
/e/	f_1	20.7	10.7	16.8	16.4
	f_2	13.9	5.8	9.1	9.6
	f_3	10.8	3.3	4.2	5.5
/i/	f_1	12.9	7.8	12.6	11.2
	f_2	13.4	4.6	6.1	7.7
	f_3	11.4	3.5	4.2	5.9
/u/	f_1	24.9	14.1	22.4	20.8
	f_2	19.6	8.8	14.1	14.3
	f_3	16.7	4.2	6.9	9.1

2.2 两种线性变换的规范效果

听觉变换后，进一步用共振峰比或共振峰差两种线性规范，都能收到提高韵母间的区分和减少发音人变异的双重规范效果。表 2 是对不同规范所作的线性判别分析的结果。

表 2 不同规范的正确判别率

单位	规范与参量	正确判别率	
		元音	发音人类别
Hz	f_2, f_3	76.7	85.0
Bark	f_2, f_3	75.8	85.8
Erb	f_2, f_3	79.2	85.0
Hz	$\lg f_2, \lg f_3$	80.5	85.0
Hz	$f_2/f_1, f_3/f_2$	87.5	42.5
Hz	$\lg(f_2/f_1)$, $\lg(f_3/f_2)$	96.7	41.7
Bark	f_2-f_1, f_3-f_2	96.7	41.7
Erb	f_2-f_1, f_3-f_2	96.7	39.2
Bark	f_1-f_0, f_2-f_0	92.5	40.0
Erb	f_1-f_0, f_2-f_0	94.2	40.0
Bark	$f_1/f_0, f_2/f_0$	93.3	60.8
Erb	$f_1/f_0, f_2/f_0$	95.0	60.8

与此有关的一个问题，是韵母的特征值及其临界距离的确定。按频谱重心理论，临界距离是两特征能否为听觉区分的界线。两特征之差 (或距离) 小于临界值时，听不出它们的差别，只有大于此值时才能听出。 Chistovich 的心理物理实验结果，临界距离为 3 到 3.5 Bark，从变换后的系列 F 差值也应能估算出临界距离。表 3 与表 4 是分别以 Bark 和 Erb 为单位的普通话单韵母的 F 差值和估算的临界距离。表中数值后括符中的 + 号表示小于临界距离， - 号表示大于临界距离。

表 3 韵母特征值的距离
(临界距离 =3.0 Bark)

韵母	F_1-F_0	F_2-F_1	F_3-F_2
/i/	1.0(+)	11.3(-)	1.9(+)
/u/	2.5(+)	3.2(-)	7.7(-)
/a/	6.2(-)	1.6(+)	5.7(-)
/e/	4.4(-)	3.0(-)	6.5(-)

表 4 韵母特征值的距离
(临界距离 =3.5 Erb)

韵母	F_1-F_0	F_2-F_1	F_3-F_2
/i/	1.6(+)	13.5(-)	2.3(+)
/u/	3.9(-)	4.0(-)	8.6(-)
/a/	8.5(-)	1.8(+)	6.7(-)
/e/	6.4(-)	3.4(-)	7.3(-)

2.3 韵母在共振峰平面上的散点图

F 差和 F 比两类规范在散点图上所显示的效果，和表 2 判别分析的结果相似，比较接近，下面示例性地给出了韵母特征值在 Erb 变换和 F 差规范听觉模型作用前后的两张散点图。图中

椭圆定量地显示三组发音人的四个韵母的声学特征值的变异或散布范围．椭圆的纵横轴长为 1.8 个标准差．

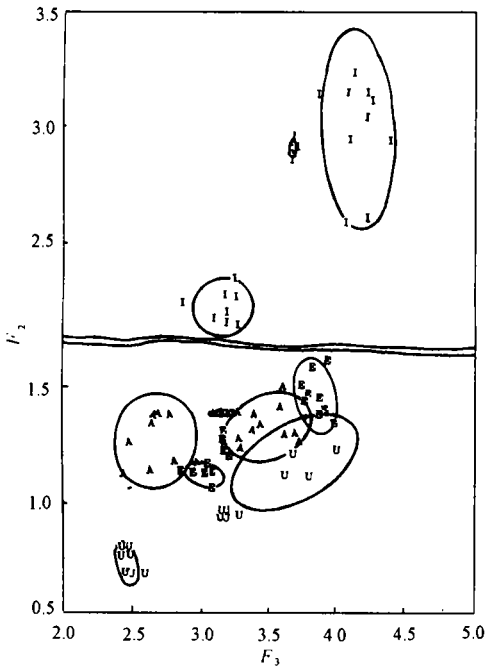

图 1　LPC 模型的韵母散点图 (单位：Hz)

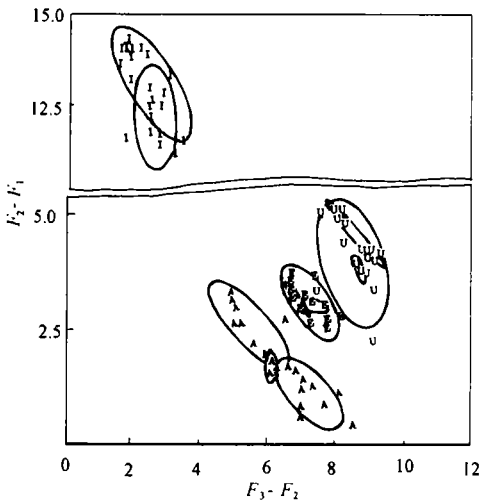

图 2　F 差模型的韵母散点图 (单位：Erb)

3 讨论

1. 不同听觉变换的效果，从表 1 中它们的离中系数比较，显然 Erb 最好，韵母特征值的变异缩小了 1—4 倍，其次为 Bark, Mel 居后．某些低频峰值的变异，Bark 还不如 Mel, 反映出两者在这一波段的特点．Deterding 曾对 Bark, Erb 的变换作过比较，没有发现明显的差别，这和他所用的指标有关，如前面表 2 所示，从识别率是很难看出两者的差异．

2. 对不同线性规范效果的比较，Miller, Syrdal 与 Gopal, 都曾有过系统的工作．Hirahara 和 Kato 的实验还表明 f_0 在规范中的作用．他们不同的结论，从表 2 的结果分析，和所用的参量有关．(1) 表 2 前四行以 f_2, f_3 为参量，判别分析表明，在线性规范前，单纯的听觉变换，无论以对数还是以 Bark 或 Erb 为单位，都不能产生规范效果，发音人类别的正确判断，和未变换前 (以 Hz 为单位) 一样，仍保持在 85% 以上．(2) 从表 2 第五行开始的线性变换表明，共振峰比和共振峰差都能收到相近的规范效果，韵母正确判别都提高到 95% 左右，而发音人类别的正确判别率都降到 40% 左右，即缩小了发音人差别引起的变异．(3) 表 2 最后四行也证实 f_0 在规范中的作用，以 $f_1 - f_0, f_2 - f_0$ 为参量，取得了同样好的规范效果，但只限于以 f_1, f_2 为参量和共振峰差的变换，否则将影响规范效果，如表 2 最后两行的结果所示．这说明，参量的选择和规范模型之间的关系还有待进一步探讨．Bladon 在讨论发音人规范的工作中就曾注意到 F_0 和 F_1 之间的密切关系．(4) 的实验结果表明，三种规范模型具有相近的规范效果，这是可以理解的，因为它们有共同的生理心理基础，即 Potter 与 Steinberg 很早指出的特定的语音在内耳底膜引起特定的空间模式，而与底膜的具体部位无关．这就是说，言语这类复合声的知觉是模式的，因而与语音

知觉常性，语音音质相对应的语声特征值不是唯一的，绝对的，而是多值的，相对的。多值性源于相对性。语声特征值的线性变换只要能表征特定的语音模式的相对关系，都能殊途同归地知觉为同一语音，保持语音知觉的常性。可以从这一观点理解 J. D. Miller 提出的“感觉参照 (sensory reference)”, Bladon 的“听觉空间位移理论”以及当前关于言语的声学变量与不变量的讨论。(4) 表 3 和表 4 的韵母临界距离的数值表明， Chistovich 的频谱重心理论所具有的普遍性。它适用于多种语言，包括汉语普通话，而且也适用于听觉单位 Erb 。普通话元音的临界距离为 3.0 Bark 或 3.5 Erb 。两数值之差显然在于 Bark 和 Erb 的带宽之差。黄国佑，曾进兴的实验结果，元音临界距离为 3.5 bark, 和虽不一致，但都在 3.0-3.5 这个可变范围内。 3. 将两张韵母共振峰平面上的散点图加以对照，清楚地显出听觉规范模型作用的效果。规范前，不同性别发音人的同一元音在图上是分散的，而且有些元音区是部分重叠的。经过规范，相同的元音从分散趋向集中，而且有些元音间原有的部分重叠也消失了。其它听觉规范模型也有相似效果，图形从略。

4 结论

通过对性别年龄不同的三类发音人的四个韵母的声学和统计分析，比较了语音识别预加工的几种听觉模型的规范效果。实验表明：

1. 三种非线性变换的听觉模型中， Erb 缩小韵母变异的效果最好， Bark 次之， Mel 较差。

2. 三种线性规范模型，无论是 F 差或 F 比或含有 F_0 的 F 差，都能收到相近的缩小韵母因发音人性别年龄不同产生的变异和提高韵母的清晰度。

3. 实验证实了 Chistovich 的 SCG 临界距离理论。普通话韵母的临界距离为 3.0 Bark 或 3.5 Er 。实验分析也表明了语音的模式性。

参 考 文 献

1 Helmholtz H L F. On the sensation of tone. 1863, (A. J. Ellis, trans. 1885)
2 Bekesy G V. Introduction, In Handbook of Sensory Physiology, 1975, ed by W. Keidel, W. Neff, **V**(1)
3 Joos M. Language. 1948, **24**(2)
4 Moore B C J, Glasberg B R. *J. Acost. Soci. Amer.*, 1983, **74**: 750—753
5 Peterson G E. *J. Speech Hear. Res*, 1961, (4)
6 Miller J D. *J. Acost. Soc. Amer.*, 1989, **85**(5)
7 Syrdal A K. *Speech Commun.*, 1985, (4)
8 Sydal A K, Gopal l H. *J. Acost. Soci. Amer.*, **79**
9 Chistovich L A, Lublinskaya V V. *Hearing Res.*, 1979, (1)
10 ILS. *Signal Technology, Inc*, 1989, **6.1**
11 Fant G. Speech Sounds and Features. 1973
12 Traunmuller H, Lacerda F. *Speech Communication*, 1987, **6**
13 Greenwood D D. *J. Acost. Soc. Amer.*, 1961, **33**
14 Deterding D H. Proceedings of 11th ICPHS (Talin). 1987
15 Hirahara T, Kato H. In Speech, Production and Linguistic Structure. 1992, ed. by Tohkure H *et al.*
16 Bladon R A W *et al. Language and Communication*. 1984, 4
17 Potter R K, Steinberg J C. *J. Acost. Soc. Amer.*, 1950, **22**
18 Disner S F. *J. Acost. Soc. Amer.*, 1980, **67**(1)
19 黄国佑，曾进兴. 第四届世界华语文教学研讨会论文集，语文分析组，世界华文教育协进会，台北： 1994
20 CSS(Complete Statistical System). Statsoft, Inc. 1987
21 SPSS(Statistical package for the Social Sciences)/Pc+ SPSS Inc. 1988, **3.1**