四、小　结

/ao 与 /ou/ 的对比分析说明，它们的根本区别在于各自的频时协变行为的不同。这个例子起码给我们提供了以下两点启示：第一，自然语音的动态特性不仅取决于频率变化的空间分布，而且受频率变化的时间分布格局的制约。因此，必须把这两方面结合起来，对它们协同变化的微观结构加以考察，才能全方位地了解它们的动态变化的规律及其相关的结构模式；第二，复合元音共振峰的动态变化既不是自始至终均匀的线性滑移，也不是两个或三个目标值之间的随意滑动，而是共振峰频率在时间轴上有定的、即遵循一定规律的滑移变化，因此，要真正认识复合元音的区别特性，仅仅知道它们的共振峰起迄频率和大致的滑移变化方向还是远远不够的。相对说来，更为重要的是应该搞清楚它们各自特有的频时协变规律及其相关的结构模式。同时，这种模式随语言或方言的不同而不同，必须对具体语言作具体分析。所以，真正了解和掌握这些特定的模式，必将有助于合成语音音质的改善和自动识别准确率的提高。

参　考　文　献

[1] Zhang Jialu, Lü Shinan, Qi Shiqian, *Journal of Chinese Linguistics*, **10**-2 (1982), 189-206.

[2] 吴宗济，«汉语普通话单音节语图册»，中国社会科学出版社，1986.

[3] Hongmo Ren, on the acoustic structure of diphthongal Syllables, UCLA Working Papers **65** (1986).

[4] Willem. J. M. Peeters, Diphthong Dynamics, Dissertation, University of Utrecht, The Netherlands, 1991.

[5] 曹剑芬、杨顺安，北京话复合元音的实验研究，«中国语文»，**6**(1984)，426—433.

27—34

普通话三合元音音节最小时间感知阈及其声学特性

祖　漪　清

（中国社会科学院语言研究所　北京　100732）

1992 年 12 月 31 日收到

本研究的实验材料取自中国社会科学院语言研究所语音数据库。库中存有 15 个男音的语音材料，共有 15×15=225 个三合元音音节。本研究的主要目的是从普通话三合元音入手，在对 15 个说话人的语音材料统计的基础上，通过对最小时间感知阈 Tlim 的测量与研究，从声学和感知的角度，给出三合元音必不可少的信息，指出多余信息。

实验结果表明，Tlim 内的共振峰变化情况可分为两类。一是动态特性，它的表现是：(*a*) △*F*1＞90%，△*F*2 约 50%；(*b*) Tlim 内至少包括 *F*1，*F*3 两个拐点中的一个；(*c*) Tlim 内包括 *F*2 变化最剧烈的部分。这四点对四个三合元音是一致的。第二类是边界条件，Tlim 受到位置和大小两方面的限制，证明其边界共振峰频率十分重要。

一、引　言

过去对于复合元音的研究，无论其出发点是从声学、生理、感知哪一个角度，所得出的结论都是：复合元音是有动程的元音，需要用目标值和过渡两方面的参数描述。

早在六十年代，Lehiste 等[1]、Holbrook 等[2]

和 Gay[3] 对美国英语的二合元音作了实验分析，将它们分为起始段、过渡段、和收尾段，并认为反映动态变化的参数最为重要。Peeters[4] 用合成的方法对荷兰语、英语、德语的几个二合元音改变起始段、过渡段、收尾段的时间结构，并进行听辨，给出了这几种语言的二合元音的动态模式。汉语普通话复合元音的研究开展得也很早，吴宗济[5-6]在声学元音图上给出了复合元音的变化过程。杨顺安、曹剑芬[7]对普通话九个二合元音的动态特性进行了归类，他们的研究结果直接为语音合成系统提供了依据，并得到了满意的合成语音音质。贺宁基[8]从语音感知的角度着手，在普通话二合元音的滑动段位置上进行切分、听辨，测量了最小时间感知阈，其结果表明：二合元音的最小时间感知阈因不同元音而不同，并与共振峰变化率存在着补偿关系。任宏谟[9]在对普通话若干二合元音和三合元音研究的基础上，以第二共振峰 F_2 为参数，提出了复合元音的截断模型，该模型可统一地描述二合元音和三合元音 F_2 的动态变化。

纵观国内外复合元音的研究状况，实验大多集中在二合元音方面，这是研究复合元音的基础。三合元音音节是汉语普通话中特有的音节，共有四个(不包括声调)：/iou,iao,uei,uai/，对它们的动态特性进行研究是十分必要的。

本研究利用中国社会科学院语言研究所语音参数库中的材料，目的是从普通话三合元音入手，在对 15 个说话人的语音材料统计的基础上，从声学和感知的角度寻找语音学的相对不变量，给出三合元音必不可少的信息，指出多余的信息。与此相关的另一个问题，是 P-center (感知中心)的问题，P-center 一词是由 Morton 提出的[10]，它反映了这样一个事实：在一个词中存在着一个位置，听者在这个位置上可感知到该词而不需听到该词的全部；Fowler[11] 认为 P-center 是和产生、感知两方面相关连的，它受到调音的控制。尽管 P-center 是从心理感知的角度提出的，我们认为可以通过它揭示其声学上的内涵，在时间域为语音的合成、识别提供参考。

二、实验方法

1. 实验材料

实验材料取自中国社会科学院语言研究所 1991 年建立的普通话单音节语音库，它们是(15 个)：忧，尤，有，又；妖；摇，咬，要；威，围，委，胃；歪，崴，外。语音库中存有 15 个男声的语音材料，因此本实验使用的材料包括 15 × 15 = 225个三合元音音节。

2. 三合元音声学模式的测量

使用语音库附带的三合元音语图(用 Kay-7800 语图仪制做)，经辨认后画出它们的基频($F0$)和振幅(A)曲线及前三个共振峰的轨迹，由使用数字化仪测量系统（用于测量语图上的基频和共振峰轨迹的装置，包括测量仪及软件)，将这些参数输入计算机，以一定格式存盘。本实验进行的各种统计分析都是调用这些数据，进行编程处理的。

3. 三合元音最小时间感知阈及其位置的测量

最小时间感知阈 Tlim 的测量分三步进行。为了简化问题，只对阴平声调的音节进行测量。

第一步为粗实验，根据梁之安[12]关于单元音，及贺宁基关于二合元音的研究，使用 Kay-7800 语图仪，采用 80ms 的窗口，对 15 个人的语音材料从头至尾，以 20ms 的步长进行切分，得到第一批刺激，我们称这种方法为扫描法。在听辨实验中，每个刺激重复三次，间隔为 2s，成为一组，组与组之间再停顿 2s，在这个时间间隔内，由受试者对该组刺激给定的书面拼音符号作出是与否的辨别(即强迫性实验)。经过第一步实验，初步判断出这些三合元音的 Tlim 是大于 80ms，还是小于 80ms。

第二步为细实验。根据粗分的结果，对 15 个人的阴平三合元音，任意抽取 5 人的材料，使用 Kay-5500 语图仪，对 Tlim 小于 80ms 的材料，窗长取 40ms，50ms，60ms；对 Tlim 大于 80ms 的材料，窗长取 90ms，100ms，110ms，120ms，甚至更大。仍使用扫描法制做刺激，听

辨结果即给出 Tlim 的值及其在音节中的位置.

第三步为精细实验. 上述步骤中的切音刺激未曾打乱,为了证实实验结果的可靠性,我们继续将两个说话人材料中窗长被定为最小时间感知阈的那条曲线的切音刺激随机打乱（共 8 组），并使每个刺激在听辨中出现 5 次(即该组刺激数目增加了 5 倍),再进行听辨实验，每个刺激仍重复三次，由受试者做出强迫性选择.

第一,二步实验的听辨受试者为 7 人,第三步实验的听辨受试者为 10 人，年龄在 20—60 岁之间,无听力疾病历史,均不了解实验目的.

三、实验结果及讨论

测得的 15 个人三合元音前三个共振峰模式见图 1，该图较好地反映了四个三合元音的性质,为本研究提供了基本参考.

由实验步骤 2 所得的结果可用立体图形表示（图 2），x 轴是归一化的时间轴，y 轴为听辨

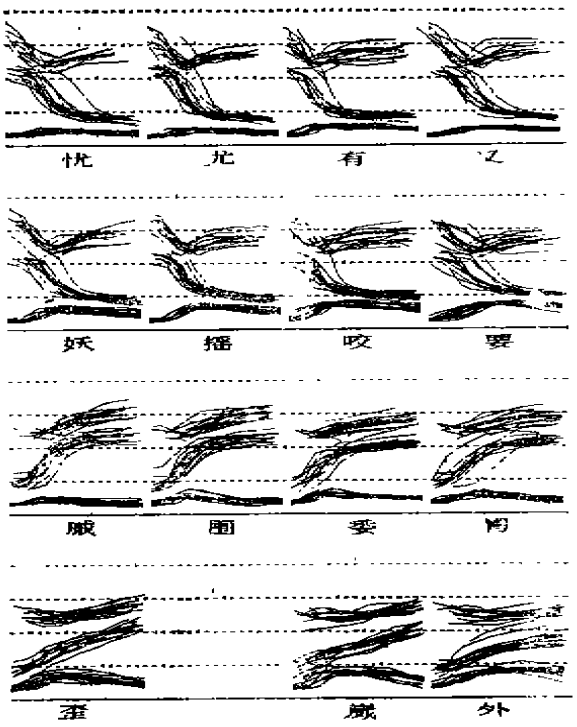


图 1 15 个发音人的三合元音的共振峰模式

表 1 五个说话人(A,B,C,D,E)三合元音 Tlim 的长度及前(Tb)、后(Te)边界归一化值

		iou	iao	uei	uai
A	Tlim.	0.121(40ms)	0.111(40ms)	0.125(40ms)	0.272(90ms)
	Tb	0.194	0.278	0.186	0.123
	Te	0.397	0.389	0.311	0.395
B	Tlim.	0.256(90ms)	0.139(50ms)	0.110(40ms)	0.357(120ms)
	Tb	0.169	0.222	0.163	0.177
	Te	0.425	0.361	0.273	0.534
C	Tlim.	0.162(50ms)	0.191(70ms)	0.220(80ms)	0.357(120ms)
	Tb	0.192	0.272	0.275	0.193
	Te	0.420	0.463	0.550	0.682
D	Tlim.	0.231(80ms)	0.201(70ms)	0.234(80ms)	0.302(110ms)
	Tb	0.172	0.229	0.233	0.163
	Te	0.461	0.486	0.467	0.521
E	Tlim.	0.150(40ms)	0.154(50ms)	0.169(60ms)	0.315(120ms)
	Tb	0.222	0.308	0.339	0.156
	Te	0.372	0.462	0.508	0.471

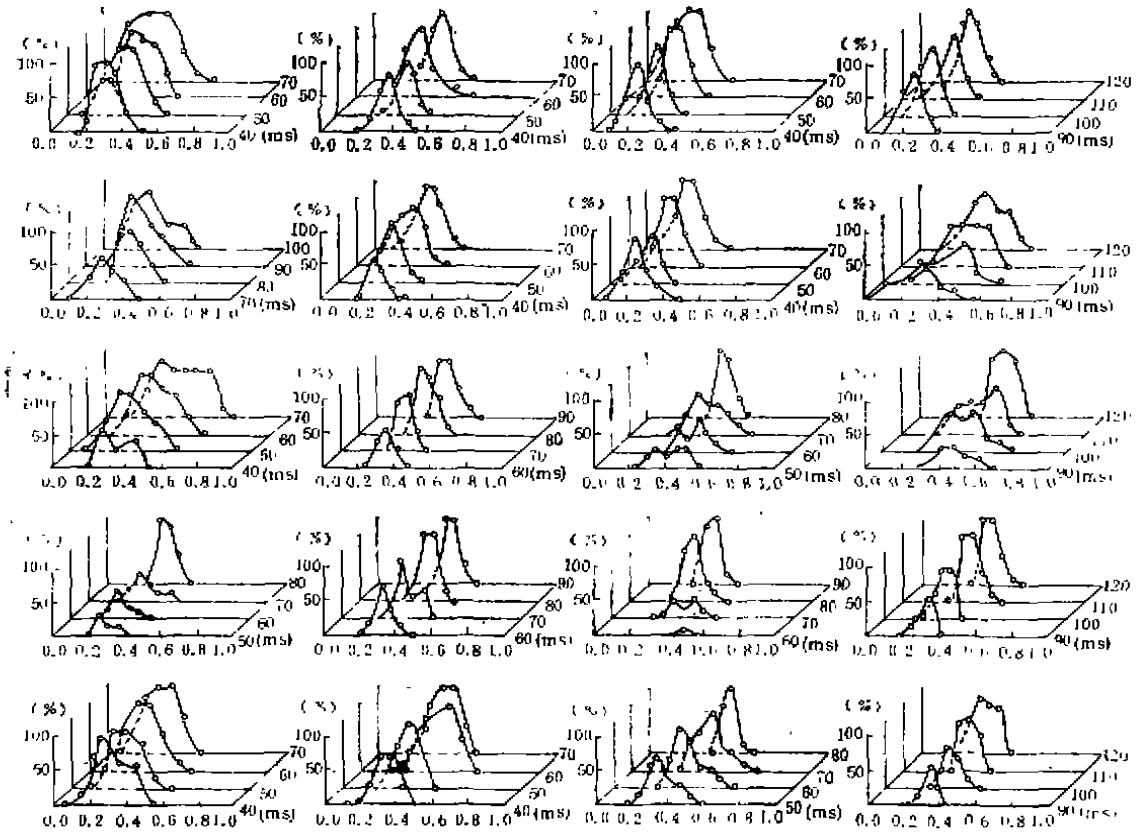
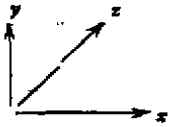


图 2　5 个发音人材料的听辨实验结果



x 轴：归一化的时间；y 轴：听辨正确率；
z 轴：感知窗口的大小。

的正确率，z 轴为时间窗口大小。由图 2 可直观地看出：最小时间感知阈 Tlim 的大小因不同三合元音而异，因人而异；其中心位置位于整个音节的前半部分：窗口越大，听辨的正确率越高，窗口增大到一定程度，听辨正确率达到稳定。

为了提高实验的可靠性，我们采听辨正确率为85%时的窗长值为最小时间感知阈 Tlim。图 3 给出 5 名发音人四类三合元音的共振峰轨迹，两条垂直线分别代表 Tlim 的前后边界，它们之间的距离大于或等于 Tlim。表 1 是这 5 个发音人三合元音的 Tlim 归一化值（括号中为毫秒数）及前后边界的归一化值。

图 4 为精细实验结果，横轴为音节的归一化时间轴，纵轴为听辨正确率。图中实线为次序打乱后的结果，虚线为次序未打乱的结果。显然，将刺激位于音节中的时间次序打乱对听辨是有影响的，使听辨正确率约下降10%—15%，但从整个曲线的形状来看，变化不大，因此可以证实由第二步实验得出的、关于最小时间感知阈的结果基本可靠。

三合元音的理想结构是：前稳段＋过渡 1＋中间准稳段＋过渡 2＋后稳段。从 15 个人的语音材料来看，前稳和后稳段的长度是不等的，但听起来它们在音色上无差异，这说明三合元音的中间部分才是最重要的。最小时间感知阈 Tlim 肯定包括了位于中间的主要元音及其向两边的过渡，否则无法感知到三合元音的色

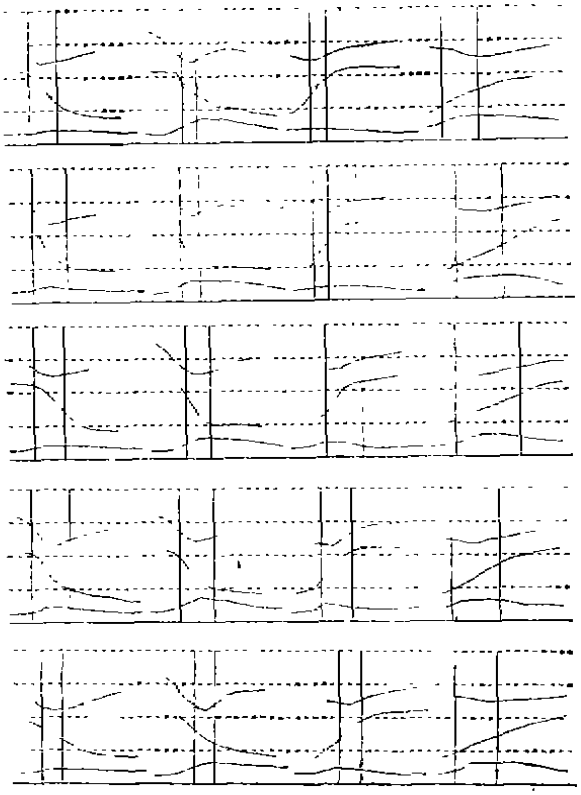


图 3　5 个发音人最小感知阈的大小及其位置
（各音节长度已进行了归一化）

彩。通过对 Tlim 这一时间轴上现象的研究，可以给出哪些是三合元音必不可少的信息。

1. Tlim 的长度

表 2 给出 5 个说话人的 Tlim，均值及标准偏差，F 检验的结果表明：四个三合元音的 Tlim 之间存在着差异（$F(3, 23) = 12.870$, $P < (0.01)$，进一步的 T 检验证明/uai/的Tlim较长，在 0.01 的显著性水平上，它与其它三合元音存在差异，而其它三个三合元音之间则无差异./uai/具有较长的 Tlim，反映了它时间分布上的特殊性，即/uai/的中间元音较长。时间长度最长的/uai/的 Tlim 值也不超过音节全长的 40%（见表 1），也就是说存在着 60% 以上的信息对感知并不重要。可以推断，如果是非零声母音节，或带有鼻尾的音节，Tlim 的长度会有所不同，但仍会存在一个小于音节长度的感知区间（关于 P-center 的研究证明了这一点）。这一事实十分重要，它对言语工程学是有一定参考价值的。

表 1 还告诉我们 Tlim 的中心位置处于整个音节的前半部分，约在前 30% 的位置上。这

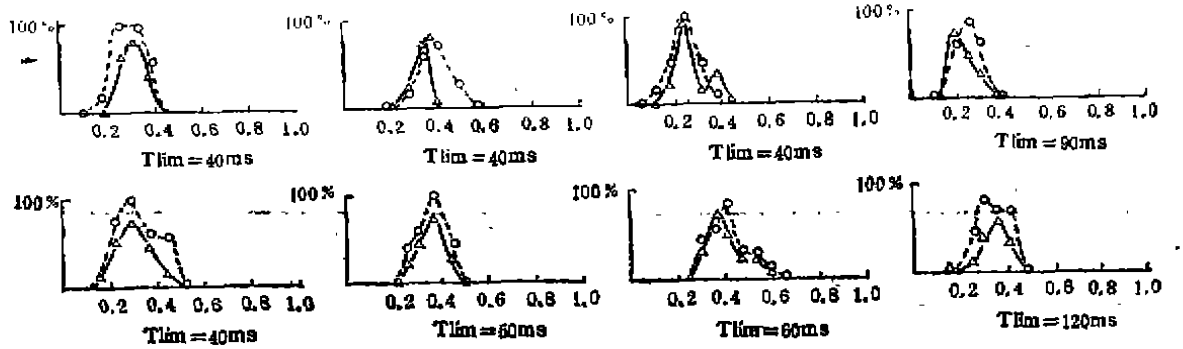


图 4　刺激打乱和未打乱的听辨实验结果（发音人 A 和 E 的）
实线：刺激打乱的；　虚线：刺激未打乱的。

表 2　5 个说话人 Tlim 归一化时长的统计分析及 F 检验结果

iou		iao		uei		uai		F value
Av	SD	Av	SD	Av	SD	Av	SD	$F(2,23) = 12.870$ ($P<0.01$)
0.186	0.054	0.159	0.037	0.172	0.055	0.321	0.037	

说明三合元音的介音占的时间较短，主要元音比较靠前，后面有一段较长的多余信息．这一结果与 P-center 的概念是相吻合的，即无须听完音节的全部，就可感知出该音节的音色，Morton 等的结果也表明 P-center 的位置在音节的前半部分，因此 Tlim 和 P-center 是有一定关系的，至少它们可以共同证明：一个音节的时间载信单元在该音节的前半部分．

2. Tlim 内共振峰的表现

从 15 个人的材料来看，三合元音区别于单元音、二合元音的基本特点是：$F1$ 有一个极大值 $F1max$，$F3$ 有一个极小值 $F3min$，这两点在时间轴上是相靠近的，但不一定重合，它们与主要元音的位置是有关的，取它们当中时间上先出现者的位置参加统计，(表 3 中的 $T1$)，T 检验的结果证明它与 Tlim 的位置无显著性差异(显著性水平为 0.01)，换句话说，Tlim 中至少包括 $F1max$，$F3min$ 中的一个(在时间上靠前的一个)，只有包括了这一点，才可反映三合元音的色彩．再看 $F2$ 的情况，取 $F2$ 最大斜率的结束位置，即 $F2$ 由较陡向较平坦变化的位置（表 5 中的 $T2$）进行统计，T 检验的结果证明在 0.01 当显著性水平上，$T2$ 与 Tlim 的中心位置 Tc 亦无差异．由此可得如下结论：在 Tlim 内，至少包括了 $F1max$，$F3min$ 当中的一个，并包括了 $F2$ 变化最剧烈的部分，由图 3 也可直观地看到这一点．

取最小时间感知阈内 $F1$，$F2$ 的起始点 ($F1b$,$F2b$) 和末点 ($F1e$,$F2e$) 频率值，以及 $F1$,$F2$(相对于整个音节)的动态范围 ($\triangle F1$，$\triangle F2$)进行统计检验，结果见表 4．四个三合元音的 Tlim 末点的 $F1$,$F2$ 存在着明显差异(显著性水平为 0.01)；$\triangle F1$ 达 90% 以上，$\triangle F2$ 约达 50%，即在Tlim 内，$F1$ 几乎完成了整个音节中的变化，$F2$ 完成了近一半的变化，F 检验的结果表明这一结论对四个三合元音都成立(对于 $\triangle F1$,$F(3,23)=1.531$,$P>0.01$，对于 $\triangle F2$,$F(3,23=0.291$,$P>0.01$)．前面讨论过/uai/的 Tlim 较长，但它在 Tlim 中，它的 $F1$,$F2$ 的变化量与其它三合元音无差异，这说明，对于/uai/，由于 $F1$,$F2$ 变化缓慢，Tlim 的长度必然加长，这正是时间和频率的补偿关系．

综上所述，Tlim 内的共振峰变化情况可分为两类．一是动态特性，它的表现是：(a) $\triangle F1>90\%$，$\triangle F2$-50%；(b) Tlim 内至少包括 $F1max$,$F3min$ 中的一个；(c) Tlim 内包括 $F2$ 变化最剧烈的部分．这几点对四个三合元音是一致的．二是边界条件，Tlim 受到位置和大小两方面的限制，证明其边界共振峰频率十分重要，四个三合元音 Tlim 边界共振峰频

表 3　5 个说话人 Tc、$T1$ 和 $T2$ 的统计结果及有关的 F 检验、T 检验的显著性水平

	iou		iao		uei		uai		F value
	Av	SD	Av	SD	Av	SD	Av	SD	
Tc	0.303	0.009	0.347	0.036	0.331	0.094	0.342	0.065	$F(3,16)=0.549$ ($P>0.01$)
$T1$	0.304	0.023	0.340	0.029	0.252	0.062	0.370	0.057	$F(3,16)=6.099$ ($P>0.05$)
$T2$	0.310	0.073	0.296	0.023	0.308	0.089	0.316	0.073	$F(3,16)=0.073$ $P>0.01$)
$Tc/T1$ T value	$T(8)=0.127$ ($P>0.01$)		$T(8)=0.357$ ($P>0.01$)		$T(8)=1.568$ ($P>0.01$)		$T(8)=0.727$ ($P>0.01$)		/
$Tc/T2$ T value	$T(8)=0.213$ ($P>0.01$)		$T(8)=2.669$ ($P>0.01$)		$T(8)=0.397$ ($P>0.01$)		$T(8)=0.727$ ($P>0.01$)		/

Tc——Tlim 的中心位置；$T1$——$F1max$,$F3min$ 先出现者的位置；$T2$——$F2$ 最大斜率结束的位置．

表 4　5 个说话人 Tlim 内声学参数的统计分析结果

		iou	iao	uei	uai	F value
F1b	Av	317.5	440.0	376.7	455.0	/
	SD	39.2	69.3	15.1	94.3	
F1e	Av	405.0	668.0	403.3	652.5	/
	SD	43.8	41.5	44.6	42.7	
F2b	Av	2082.5	1896.0	1140.0	1015.0	/
	SD	157.6	153.9	70.4	108.4	
F2e	Av	1375.0	1188.0	1950.0	1655.0	/
	SD	356.5	99.6	193.0	145.3	
$\triangle F1$	Av	0.965	0.972	0.920	0.997	$F(3,23)=1.531$
	SD	0.062	0.063	0.110	0.008	($P>0.01$)
$\triangle F2$	Av	0.513	0.521	0.458	0.472	$F(3,23)=0.291$
	SD	0.195	0.169	0.119	0.054	($P>0.01$)

率各不相同. Gay[3] 认为复合元音中对感知起关键作用的是起点频率和共振峰变化率，贺宁基[8]关于汉语普通话二合元音 Tlim 的研究支持了 Gay 的观点. Bladon[13] 在他的实验中切掉英语二合元音的过渡部分，只保留稳定部分，或切掉稳定部分，只保留过渡部分，进行听辨，他的结果是端点共振峰频率更为关键. 我们通过对三合元音的研究，认为 Gay 和 blandon 讨论的是问题的两个方面，就 Tlim 来说，端点共振峰频率更为重要；对整个音节而言，是变化最重要，即 Tlim 的信息最主要. 在 Tlim 内，端点频率相当于边界条件，在它的规定下，共振峰频率以一定的形式实现各三合元音的个性，四个三合元音的动态特性是它们的共性（区别于单元音和二合元音），边界条件则代表了它们的个性.

3. 利用知识进行的三合元音的识别

根据以上三合元音的分析结果，利用存放的三合元音共振峰参数对 15 个人的三合元音进行识别，可选用如下参数

(1) $Mf2 = F2B - F2E$，判别介音是/i/还是 /u/；(2) $RFhal$（音节中点 $F2$ 的变化率），判别是/uei/还是/uai/；(3)$F1max$，判别是/iou/还是/iao/.

用以上参数对语音库中 225 个三合元音进行识别，识别率为 98%. 我们可以看出，这些参数都不是反映音节两端变化较小的部位的，而是和 Tlim 相关联的，这部分实验也是对最小时间感知阈重要性的论证.

四、结　论

1. 零声母三合元音音节的前半部分存在着一个小于音节长度 40% 的感知阈，或者说零声母三合元音有 60% 以上的信息对感知并不重要；

2. Tlim 的大小与 F2 的变化率成反比；

3. 三合元音区别于其它元音的主要特征是 F1 有一个极大值，F3 有一个极小值；

4. 三合元音 Tlim 携带了三合元音的主要信息，它的端点频率及其由它规定的动态变化构成了三合元音的个性.

参 考 文 献

[1] Lehiste J., and Peterson G., *J. Acous. Soc. Am.*, **31** (1961), 268—277.

[2] Holbrook A., and Faibanke G., *J. Speech Hearing Res.* 5-1(1962), 38—58.

[3] Gay T., *J. Acous. soc. Am.*, **44**-6(1968).

[4] Willem J. M. Peeters, "Diphthong dynamics-A cross-linguistic perceptual analysis of tem poral patterns in Dutch, English and German", Ph. D. dissertation, 1991, University of Utrecht.

[5] 吴宗济等，《普通话单音节语图参考册》，中国社会科学出版社，1986.

[6] 吴宗济，曹剑芬，中国语文，4(1979)，314—320.

[7] 杨顺安，曹剑芬，语言研究，1(1984)，15—22.
[8] 贺宁基，“北京话二合元音感知中的时间因素”，《北京语言实验录》，北京大学出版社，1985.
[9] Ren Hongmo, UCLA working papers in phonetics, 65(1986).
[10] Morton J., Marcus S., and Frankish C., *Psychological Review*, 83(1976), 405—448.
[11] Fowler C. A., *Perception and Psychophysics*, 25 (1979), 375—388.
[12] 梁之安，声学学报，2(1965)，20—23.
[13] Bladon A., *Speech Communication*, 4(1985), 145—154.

34-38

人工神经元网络，语音识别 8
声图象识别

自组织人工神经元网络用于声图象识别的研究

TP 391.4
TP18

桑恩方　乔晓宇　季　瑞
（哈尔滨船舶工程学院水声研究所 哈尔滨 150001）

1992 年 11 月 30 日收到

本文通过实验研究了自组织人工神经元网络用于声图象识别的步骤和方法，讨论了在水声、超声医学等声图象识别中所遇到的一些关键性技术问题.

一、引　言

八十年代以来，人工神经元网络的研究取得了新的重要进展. 由于它们的并行性、分布式存储、自学习、自组织的结构特点，对传统人工智能有了较大突破. 因而在许多领域已展现了广阔的应用前景.

在声学领域中，随着海洋开发、海底沉物探测、智能机器人声视觉及超声医学等事业的发展、人们已发展了多种二维和三维高分辨率声成象技术. 随之必然地提出如何根据声图象进行被探测物体的自动识别和理解的任务要求. 借助人工神经元网络来较好地完成这一任务，是本文研究的主要目的.

已有许多重要的神经网络模型被提出.[1—3] 由 Kohonen 提出的自组织算法模型[4,7,8] 是其中代表研究之一. 由于它所模拟的是人脑神经对外界刺激具有自动排列和顺次响应的功能，因而具有很强的分类性. 已用这种模型设计出矢量量化器. 用于语音识别和图象编码.

完成计算机“识别”和“理解”的基础是实现一个自动分类器. 本文主要研究如何根据声图象的几何特征进行自动分类，从而为目标的进一步识别和理解打下基础.

二、自组织人工神经元网络模型

以 Kohonen 网络算法为代表的自组织算法是一种无“教师”学习的方法. 输出可看成是输入样本特征的一种表达. 它是一个双层网络，其输出节点是在平面上按顺序排列的. 其算法为：

若网络有 N 个输入节点，M 个输出节点，给出一组初权向量 W_{ij} 后，对于在时刻 t 给定的一个样本 $X_i(t)$，计算距离

$$d_j = \sum_{i=0}^{N-1}[X_i(t) - W_{ij}(t)]^2$$

$$0 \leqslant j \leqslant M-1 \quad 0 \leqslant i \leqslant N-1 \qquad (1)$$

选择距离最小的输出节点为响应节点 j^*，然后修正 j^* 及其领域内输出节点连接的权值：

$$W_{ij}(t+1) = W_{ij}(t) + k(t)[X_i(t) - W_{ij}(t)] \qquad (2)$$

其中 $0 \leqslant j < M^- - 1$，$0 < i < N-1$；$0 <$