综述·

听觉场景分析及其评价①

王　净　　杨玉芳

中国科学院心理研究所（北京　100101）

[摘要]听觉场景分析部分地借助于格式塔原则来研究听觉组织加工过程。它包括初级分析和图式加工过程。初级分析着重研究序列整合与同时性整合；而图式加工涉及到注意与知识的作用以及言语知觉的特殊性等。场景分析能够较好地说明简单音和复合音的知觉组织过程，但尚不能对言语加工进行全面地解释，其理论仍然存在一定局限性。

关键词　场景分析，图式，初级分析，整合，分组
分类号　59.831

1 听觉场景分析

听觉场景分析（auditory scene analysis）是用来研究听觉系统如何对外界刺激进行组织与加工的。其任务有两个：一是找出那些能够使声谱成份组合到一起或使它们分离成独立的听觉流或表象的声学特征；二是研究听觉分组的方法[1]。场景分析包含两个阶段[2]，一是以格式塔原则为基础的初级分析，它把不同感觉元素分配到相应组中；另一阶段是图式加工，它可以对知觉组织进行验证和修复。这两个阶段分别对应于自下而上和自上而下两个加工过程。

1.1 初级分析

初级分析过程是先天的，无需有意注意参与[2]。其策略是：先把听觉信号分割成许多独立的单元，这些单元与声谱中特定时域和频域相对应。然后，对这些单元进行分组或分离。分组是指听觉系统把某些具有相似特征或时间接近的音知觉为一个流，使之从复杂的环境声中突出出来。分离则是从复杂环境声中辨别出声音的不同来源或区分不同声音。分离和分组是一对统一的概念，如果出现了分组，也就意味着流与流之间产生了分离。初级分析包括序列整合和同时性整合。前者把在不同时间内顺序出现的谱成份纳入一个知觉流，以便计算环境中声音的序列特性。而后者则把同时出现的成份分开，将它们放入不同的流中。

1.1.1 序列整合　序列分组中的流形成遵循接近性和相似性原则，时间或频率接近的音将被分成一组。序列整合有两种形式，一是对由两个音高不同的简单音交替出现构成的序列的整合，听者会把这个序列听成两个来源不同、分别由低音和高音构成的流。另一种形式是对由频率关系较为复杂的音所构成的序列进行整合。把两组频率关系变化的音，按一定顺序排列，构成一个序列，该序列能否产生曲调取决于音之间的频率关系。当两组音的频

① 本文于 1998-02-11 收到。

域相同时，曲调将消失；但如果它们分别在两个没有重叠的频率范围内，则曲调便被听成一个独立的流。

影响流分离最重要的因素有两个：一是交替速度；二是两个交替出现音之间的频率差。序列速度取决于音之间的时间间隔。Bregman 等[3]认为，产生分组效应的时间间隔为 35ms。Darwin 等[4]发现，当单个谐波提前或滞后 32ms 时，这个谐波就会与其他谐波产生分离。Dai 等[5]甚至认为，听者可以按照特定任务要求，去调节时间整合间隔。但究竟是相同频率音之间的间隔重要，还是两个不同频率音之间的时间间隔重要？这个问题还有待进一步研究。

序列分组的影响因素还有基频、时间接近性、谱形状、强度和空间位置等。这些因素在分组中有竞争也有合作。如果所有因素对分组都有促进作用，则分组将被加强。例如，空间差异与其它因素联合起来时，其作用最强；又如，仅响度不同的两个音可能不会产生分离，但如果加入其它差异，则响度就可能起重要作用。

1.1.2 同时性整合　声音中的谐波在频域中呈线性分布，而在基底膜上激活的相应位置则呈对数分布。在对数频率单元中，低次谐波之间相距较远，而高次谐波则相距较近。因此，谐波捕获有如下规律：(1) 复合音中低次谐波比高次谐波更容易被捕获；(2) 包含奇次谐波的谱成份比包含连续谐波的谱成份更容易被捕获；(3) 相邻谐波被剔除的谐波易被捕获。因此谐波之间的频率相差越大，这些谐波越容易从复合音中被捕获[2]。Duifhuis 等[6]发现，在音高知觉中，听觉系统能够把额外声从复合音中剔除，其作用原理如同筛子，对信息进行过滤，他们把这种加工机制称作“谐波筛”(harmonic sieve)。那些与基频很接近的谐波可以通过筛子；而其它额外声则无法通过。这种效应在低信噪比情况下更为突出。谐波筛是流形成的机制之一，其工作方式不是全或无 (all - or - none) 的，而是渐进的[7]。谐波在失调比例很小 (3 % - 8 %) 的情况下，谐波仍然能够通过谐波筛。谐波筛只对那些可分辨的谐波起作用，它不能完全把额外音从元音的第一共振峰频域中剔除，因为两者之间可以产生部分整合[8]。

调频 (FM) 和调幅 (AM) 对同时性整合同样产生影响。FM 中，听觉系统使用 FM 范围差异对同时呈现的音进行知觉分离[9]。FM 有两种，一是等差调制，它把原来的每个谐波都加上相同的频率。通过这种变化之后，它们之间的谐波关系消失，从而使音的不同成份产生分离。另一种是等比调制，即每个谐波被乘以相同的整数，这种处理之后，其谐波关系没有改变，但谐波之间的间距被扩大了。AM 使不同谱位置上振幅产生变化。这种变化以及它们的出现时差和消失时差都对流分离产生影响。振幅的同步变化可以使谱产生分离，这与神经活动特征相一致。对应于不同谱位置的神经元的同步活动，保持时间很短。在频谱图上的每一段 (segment) 内，相对应的神经元同步活动，而段与段之间则不同步。音的识别就是通过对这些段中同步性变化的觉察而实现的。

空间位置信息和频率信息的加工可能是独立的。生理学研究发现，猫的主要听觉皮层受损时，它对某个频率声音的空间位置判断能力将丧失，但仍能够判其它频率声音的空间位置。在人脑中也存在类似现象。另外，人可以同时听到不同位置上频率不同的两个纯音，两者不会融合。例如，在 250 - 4000Hz 内，最大频率差超过 7 %时，便不会产生融合。虽然听者把左右耳声音听成两个独立的音，但当双耳听到的音在频率上接近时，便出现双耳整合。

序列整合与同时性整合的差别是显而易见的。然而，两者都涉及到声音的最基本特征：时域和频域，所以，它们不是孤立的。在复杂的声音知觉中，两者共同参与，相互影响。

1.2 以图式为基础的知觉组织

听者把环境中特定的声音信号，如言语、音乐以及其它熟悉的声音等存入记忆中，形成认知单元，这些认知单元就是认知图式。当听觉系统获得的信息模式与图式相同时，图式将被激活，并且通过图式对模式的其余部分进行推测。图式还可以被与其相关联的其它图式激活。图式加工是一个自上而下的加工过程，知识和注意在这一加工过程中起着重要作用。

1.2.1 注意与知识的作用　图式加工需要对信息进行选择，它与注意关系密切。有意注意可以控制图式，只要任务要求注意参与，图式就会出现，即产生以图式为基础的分离。听觉系统可以利用频率线索把注意集中于一个特定的频率范围内。注意参与的加工过程可以很容易地把一个流分解，但不能对流之间的信息进行整合，因为注意只能指向一个流[10]。图式形成过程就是获得关于刺激知识的过程。利用获得的知识，听者可以对刺激的变化趋势进行预测。当我们听一个重复的音时，就会掌握它的规则，形成图式。这些规则知识使我们心理上做好准备，把这一序列整合到连续的心理表征之中。由于规则可以很容易地使注意集中，所以它对流形态的影响很大。轨迹规则效应在有记忆参与的任务中最明显，以轨迹为基础的组织效果随刺激呈现次数增加而增加。听觉模式识别存在着图式成份的激活，这种激活受到时间规则和其它形式规则的影响。

1.2.2 语音的加工　对言语声的组织同样有两种形式：序列整合和同时性整合。序列整合过程把按顺序出现的词（或元音）的各部分整合到一起，这是词识别的基础。产生这种整合的前提条件是相邻部分的声学特性（如：音高、共振峰以及基频等）具有连续性，或相差不大。语音的流分离与非语音的流分离一样，也受到序列速度的影响，速度越快分离程度也越大。另一种形式的组织是同时性整合，一般情况下，我们所面对的往往不是一个孤立的声音，而我们所能够意识到的可能只是其中的一个声音。这是因为听觉系统能够对众多的声音进行同时性整合，从而形成一个知觉流。同时性整合过程中，基频（或音高）是一个重要线索，差异越大，越容易分离。共振峰也表现出相同的作用规律。此外，声音的空间位置在同时性整合中也起一定的作用，不同位置的声音很容易被区分。

与简单音模式相比，言语信号很不稳定，不仅它的基频随时间变化而变化，而且还包括与基频同步变化的口腔共振声。因此，语音加工和言语知觉过程更为复杂，加工过程中，既有初级分析，又存在着图式加工。一些初级分析模型只适用于简单元音或与元音相类似的音，而无法解释象“鸡尾酒会效应”等复杂过程；如果用图式加工来解释则非常容易，即熟悉的声音或内容激活了记忆中相应的图式。

1.3 初级分析与图式加工之间的关系

初级分析和图式加工在功能上不是独立的，可能存在相互竞争[11]。Bregman[2]认为，(1) 图式加工在初级分析之后进行；(2) 通过训练和学习，信息就可以用图式表征；(3) 图式不参与听觉分析中的流形成过程；(4) 图式描述的是知觉元素的典型特征，不论辅音、元音、音节或其它言语成份都是如此。以格式塔原则为基础的听觉分组是低水平加工的结果。输入到听觉系统中的信息是杂乱的，这些无序波动将在听觉信息加工的后期阶段被最大限度地减弱。

两个过程对知觉的影响不同。初级分析过程把感觉信息分离，而图式加工过程则对信息进行选择，而不是将它从混合音中移走。初级分析中的分组是对称的，当它把高音和低音分离时，就形成两个独立的流。同样，也可以把两个来自不同空间位置的音分离开，判断出一个音来自左边，另一个来自右边。图式加工过程则没有这种对称性。在一个混杂环境中，我们可以很容易地听出自己的名字，但并不能辨认出在名字出现时，背景音是什么。因此，可以用分离的不对称性来确定是否出现了以图式为基础的分离。初级分析与图式加工的时间范围也不同，图式加工过程所涉及的时间范围比初级分析的时间范围大。

由此可见，初级分析和图式加工是两个不同阶段，前者是场景分析的第一阶段，当它不能解释言语组织过程时，就需要用图式来说明。把通过初级分析而产生的分离与图式加工结合起来，就可以避免或减少因初级分析和言语的熟悉性而产生的错误。两个系统所加工的感觉信息相同，但两者的加工难度不同，初级分析比图式加工更困难一些。

2 场景分析的总体评价

传统的听觉理论主要从生理学角度解释人的听觉过程，例如：地点学说、行波学说、齐射学说等，而场景分析从心理模型角度，把格式塔原则和图式过程应用于听觉信息加工；把听觉组织过程看成一个具有层次性的加工过程，丰富了听觉组织的理论。场景分析还用生态学的观点来分析听觉组织过程，从而增加了理论的外部效度，使之更加易于实际应用。然而，在某些理论与方法上，仍然存在一定局限性。

2.1 格式塔原则的局限性

当分组原则相冲突时，如果未产生融合，则格式塔原则是有效的。相反，如果所有的格式塔原则对分离或分组都有控制作用，并且产生知觉融合，那么非格式塔原则也必须参与，知觉组织才能顺利完成[12]。由于每个人所发出的音都有其特定性，在言语知觉中，不是把那些简单的声音模式分组，而是要识别发音器官所发出的复杂声音；此外，还必须借助于已有的言语图式或其它知觉经验才能理解其含义。

2.2 关于图式的作用

Bregman 只总结了那些有利于场景分析的研究结果，而忽略了一些反面的研究结果[12]。格式塔理论在实验中很容易体现出来，而图式成份的作用则难以说明。场景分析把图式过程看成是万能的，不论何种条件下，一切格式塔原则所不能解释的现象都用它来解释。这种过分夸大图式作用的观点忽略了图式本身的重要特性。另外，图式的层次性、大小及相互关系等特征还有待于进一步研究。

图式的来源直接影响到对图式作用的认识。与初级分析相比，图式是学习获得的[2]。但 Eimas 等[13]发现，3、4 个月的幼儿可以对不熟悉的信号成份进行空间的或谱的整合，以便区分它们，尽管没有对他们进行过这方面的训练。乔姆斯基认为，要确认对儿童言语的强化是极端困难的，甚至是不可能的。所以，图式也并非完全都是后天习得的。

另一个问题是，言语组织是否能够持续到图式加工阶段？在图式驱动过程中，需要去检查是否存在错误的组织。然而，言语信号保持时间很短，在语音编码的快速期内，200ms 之后，音节的感觉痕迹就消失了。甚至只能保持 100ms 或更短。如果由于听觉痕迹的易变性，使得语音机制和其它听觉机制必须同时存在，那么这就与场景分析中初级分析和图式加工的序列结构系统相矛盾。

参考文献

[1] Grose J H, Hall J W, Mendoza L. Perceptual organization in a co-modulation masking release interference paradigm: Exploring the role of amplitude modulation, frequency modulation, and harmonicity. *The Journal of the Acoustical Society of America*, 1995, 97: 3064-3071.

[2] Bregman A S. Auditory Scene Analysis: The Perceptual Organization of sound. Cambridge, MA: MIT Press, 1990.

[3] Bregman A S. Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, 4: 380-387.

[4] Darwin C J, Pattison H et al. Vowel quality changes produced by surrounding tone sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 1989, 45: 333-342.

[5] Dai H P, Wright B A. Detecting signals of unexpected or uncertain durations. *The Journal of the Acoustical Society of America*, 1995, 98: 798-806。

[6] Duifhuis H, Willems L F. Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, 1982, 71: 1568-1580.

[7] Moore B C J, Glasberg B R. Thresholds for hearing mistuned partials in the pitch of complexes. *The Journal of the Acoustical Society of America*, 1986, 80: 479-483.

[8] Roberts B, Moore B C J. The influence of extraneous sounds on the perceptual estimation of 1st-formant frequency in vowels. *The Journal of the Acoustical Society of America*, 1990, 88: 2571-2583.

[9] Plack C J, Carlyon R P. The detection of differences in the depth of frequency modulation. *The Journal of the Acoustical Society of America*, 1994, 96: 115-125.

[10] Botte M C, Drake C et al. Perceptual attention of nonfocused auditory streams. *Perception & Psychophysics*, 1997, 59: 419-425.

[11] Ciocca V, Bregman A S et al. The phonetic integration of speech and non-speech sounds: Effects of perceived location. *Quarterly Journal of Experimental Psychology*, 1992, 44A: 577-593.

[12] Remez R E, Rubin P E, Berns S M et al. On the perceptual organization of speech. *Psychological Review*, 1994, 101: 129-156.

[13] Eimas P D, Miller J L. Organization in the perception of speech by young infants. *Psychological Science*, 1992, 3: 340-345.