

第 1 章

语音加工研究

1.1 语音加工的研究问题

近期的语音加工研究关注不同语音感知的现象和话题。其中，由于一语或母语语音加工通常相较于二语语音加工更加稳定且准确，因此一语和二语的语音加工研究有着不同的侧重点。由于文献中一语的语音加工研究涉及更加丰富的话题，本章介绍的一语研究主要涉及四类研究问题：（1）语音感知加工是否存在个体差异？这些个体差异由何而来？（2）语音感知和产出是否相关？（3）汉语声调加工的感知过程中，不同声学特征的感知比重有何差异？（4）新生儿、幼儿的语音加工机制如何发展？有何特点？老年人的语音感知如何衰退？语音感知的老化受何种因素的影响？特殊人群（如听障人士、人工耳蜗植入者）的语音感知有哪些特征？受何种因素的影响？

二语语音加工研究中，不同类型语言双语者的二语语音加工研究已获得较为长足的发展。本章介绍的二语研究主要涉及三类研究问题：（1）汉语母语者的二语语音感知的特征如何？其感知特征与什么因素有关？（2）汉语二语者对汉语语音的感知有何特征？（3）二语语音的习得如何影响一语语音的感知？受到何种因素的影响？

1.2 语音加工的研究方法

语音作为语言的主要承载者，具备不同于非语音声音的特殊声学特征。不同的语音声学特征携带着不同的语音信息。在考察语音加工过程

中，研究者通常需要在行为或认知神经科学实验中操纵所关注语音特征的相关声学指标，从而考察听者在语音加工过程中的行为表现特征以及相应的神经加工机制。

本小节首先介绍语音中的主要特征和相关声学指标，并对语音加工相关研究中常用的行为、认知神经科学实验方法及其常见的数据指标进行简要介绍。

1.2.1 语音中的特征和指标

语音中包含音段、超音段特征，其中音段特征主要指元音、辅音等构成音节（syllable）的语音成分；超音段特征是指音节以上的声学特征，如音高（pitch）、重音（stress）、语速等。

1. 音段声学特征

音段特征以元音、辅音等音位为主。元音指语言中发音时气流不受阻碍的音位（phoneme），如 /i e o/ 等，由不同的舌位下声带振动发音产生（Mitterer & Cutler, 2006）。元音的响度较大，是音节的主要构成部分。辅音主要指语言中发音时气流受到一定程度阻碍的音位，如 /p k t/ 等，能够构成音节的头部和尾部。辅音由口腔、咽腔、鼻腔、唇、齿、齿龈、硬腭、软腭和小舌等不同的发音部位，以不同的发音方式（如爆破、摩擦等）发音产生，形成不同辅音上各不相同的语音和声学特征（如爆破音声波中包含持阻的静音和爆破时声波的突然放大，而摩擦音通常呈现出与噪音相似的声波和图谱等）。

元音的主要声学特征为第一、第二共振峰（formants）。共振峰指语音的声波中能量集中的频率带，在语音的三维频谱图（spectrogram）中呈现为贯穿元音主要发音频谱的深色粗频率带（如图 1-1 所示）（Ladefoged & Johnson, 2001）。在元音的频谱图中，自下而上会出现第一共振峰（first formant, F1）、第二共振峰（second formant, F2）和第三共振峰（third formant, F3）。元音的共振峰特征能够集中反映元音的音质：第一共振峰与元音舌位高低成反比（舌位越高，F1 越低）；第二共振峰与元音的发音舌位前后成正比（舌位越前，F2 越高）；第三共振

峰则与元音的圆唇、卷舌等有关。

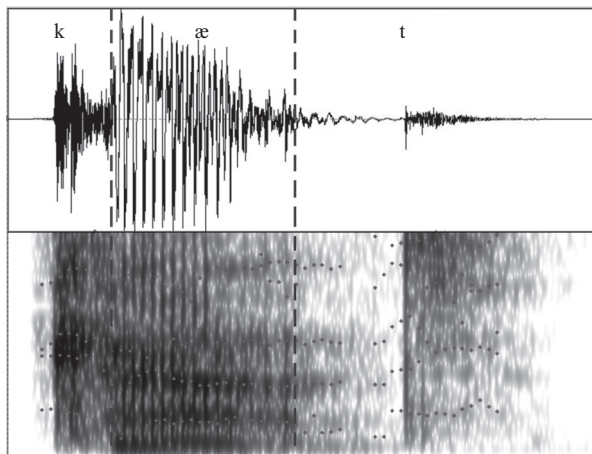


图 1-1 单词“CAT”的声波图(上)和三维频谱图(下)¹

注：声波图中声波范围最大、频谱图中颜色较深的部分为元音 /æ/；声波图中的虚线为三个音的大致分界线；三维频谱图中黑点线所示部分为声波中的共振峰，自下而上为 F1、F2、F3、F4。

音长 (duration) 也是元音研究中常见的指标。一方面，部分元音区别对除了存在音质上的差异，还在元音的音长上存在差异，如英语中的 /i/ 和 /ɪ/ 区别对中，/i/ 音相比 /ɪ/ 音的舌位更高、更前，因而音质上 F1 更低且 F2 更高；在实际语音产出中，/i/ 音的音长通常长于 /ɪ/ 音。因而，元音研究中也常以音长为考察指标。

辅音研究中，爆破音的噪音起始时间 (voice onset time, VOT) (Mitterer & Cutler, 2006; Shadle, 2006) 是一项重要声学指标。爆破音的噪音起始时间指的是发音从除阻爆破到声带开始振动发音的时间，计算公式为：T (声带振动) - T (爆破音除阻)。对于清辅音 (如英语中的 /p t k/)，爆破音除阻早于声带开始振动的时间，VOT 为正值，且送气音 /p^h t^h k^h/ 的 VOT 大于不送气音 /p t k/ 的 VOT (后者的 VOT 通常在 20 毫秒以下)。而对于浊辅音 (如英语中的 /b d g/)，爆破音除阻前

1 单词语音发音取自剑桥英语词典 (Cambridge Dictionary) 网站，声波图取自语音处理软件 Praat，相关简介可见 1.2.1 语音声学特征分析工具。

声带就已经开始振动，因而 VOT 为负值（如图 1-2 所示）。不同语言中，虽然许多爆破音属于同一音位，但它们通常存在 VOT 差异，例如，英语中的 /p/ 的 VOT 长于法语中的词首 /p/ 音。除了 VOT 的差异，研究者也会根据不同语言中的特殊的语音差异进行探究，如 Idemaru et al. (2012) 以日语中不同持阻时长（即爆破音开始持阻到除阻间的时间）的词中爆破音为例考察被试的语音加工特征。

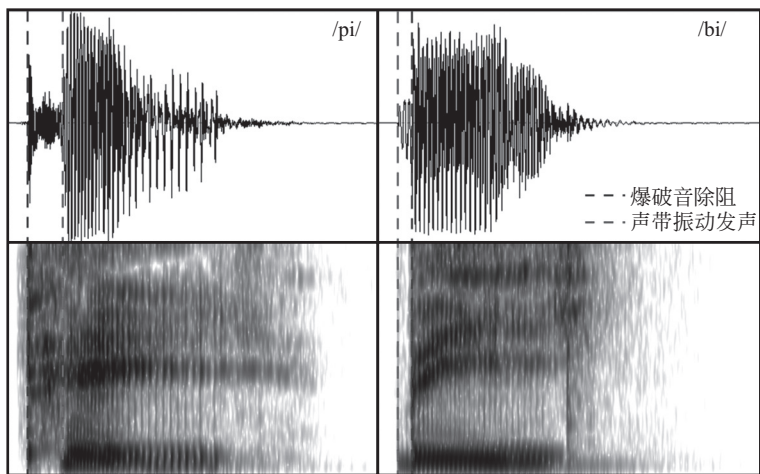


图 1-2 英语中的 /p/ 和 /b/ 的发音 VOT 示意图¹

注：虚线表示 p 和 b 声除阻，点虚线表示声带开始振动发声，虚线中间的区域即为 VOT 计算区域。

2. 超音段声学特征

超音段特征主要包含了音高、重读、音调（intonation）、音长等特征。其中，音高指的是声音被听者感知时，从低沉到尖锐的听觉差异（Gandour, 1978）。音高的高低取决于声带振动的快慢，声带每秒振动次数越多、声带振动越快，则音高越高。由于喉部生理构造的差异，男性的音高通常低于女性，听感上更为低沉。音高本身是一项非语言学的声学特征。而声调（tone）指的是音高在语言中使用的特殊形式（Gandour, 1978），在汉语普通话、粤语等声调语言（tonal language）

1 语音音频来自剑桥英语词典网站，声波图和频谱图取自 Praat 软件。

中有着区分文字和语义的重要作用，如汉语中由同一音位不同声调构成的不同汉字“妈、麻、马、骂”。音高和声调通常以语音的基频（fundamental frequency, F0）为主要的声学指标。

3. 语音的声学特征分析工具

在语音加工研究中，研究者通常需要对被试感知或产出的目标语音特征进行声学分析和加工。由 Boersma & Weenink（2018）开发的计算机软件 Praat¹，是进行语音声学特征分析和声学标注的一款功能强大且较为简洁易用的计算机软件。Praat 能够将语音音频文件以声波图和三位频谱图的形式呈现在电脑上，支持研究者进行共振峰、音高、响度、音长等语音声学特征的分析，以及指标数据的标注和提取，并能够对语音进行简单的裁剪拼接等处理。借助 Praat 的声学分析和处理功能，研究者还可以利用 Praat 脚本进行语音研究实验程序的编写。图 1-3 为 Praat 程序界面。

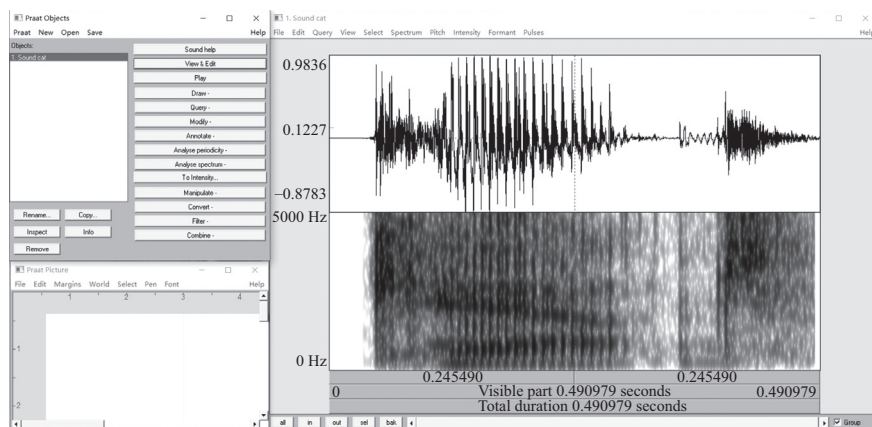


图 1-3 Praat 程序界面

注：Praat 软件基础功能窗口为左侧的 Praat Objects 和 Praat Picture，用户可以从 Praat Objects 窗口读取语音音频文件，可以在右侧的 Edit 窗口查看语音文件的声波图和频谱图，并对语音音频进行分析、编辑等操作，也可以将语音的声波和频谱图输出到 Praat Picture 中查看。

1 来自 Praat 软件官方网站。

1.2.2 行为实验方法：语音感知和听辨任务

1. 语音辨别任务

语音辨别任务 (discrimination task) 要求被试分辨刺激语音是否为同一音位或声调的语音。在该任务中, 被试需要先后听两个音 (AX 语音辨别任务) 或三个音 (ABX 语音辨别任务), 听完刺激音后, 判断刺激音是否一致或多大程度相似, 或三个音中哪两个音更相似。在这类任务中, 研究者可能操纵刺激音的某些声学特征, 以检验被试能否感知其中的差异。被试对不同类型语音的辨别数量和判断相似的百分比为常用的语音感知指标。

2. 语音识别任务

语音识别任务 (identification task) 是另一类常用的语音感知任务。在这类任务中, 被试在每个试次中听一个刺激音, 并从视觉呈现的备选音标中选择所听到的刺激音。这类任务通常用于测量语音的范畴化感知 (categorical perception), 研究者常针对对立音连续体 (continuum) 生成在特定声学特征 (如 VOT) 上等量递进变化的语音, 主要感知指标为被试在不同步数 (即从对立音一端到另一端之间的递进次数) 上选择目标音的数量和百分比, 及其形成的范畴化感知变化的斜率。

3. oddball 实验范式

oddball 实验范式 (oddball paradigm) 是一种通过在连续的标准刺激中插入偏差刺激, 要求被试在感知到偏差刺激时迅速做出反应的实验范式, 能够检验被试对偏差刺激感知的敏感性 (Carbajal & Malmierca, 2020; Kropotov, 2009)。在语音感知研究中, 通常采用听觉 oddball 实验范式检验被试在连续重复语音刺激中对偏差刺激的反应速度和正确率 (如 Zhao et al., 2022)。

1.2.3 行为实验方法：眼动追踪任务

眼动追踪技术能够收集被试在视觉呈现的语言加工过程中的注视、扫视等眼动行为，反应被试语言加工过程的特征和加工难度等。在语音加工研究中，研究者通过操纵视觉呈现材料中的语音特征，观察被试对目标材料的注视情况，能够推断被试的语音加工特征和过程。例如，刘雪丽和倪传斌（2022）通过视觉情境范式（visual world paradigm, VWP），观察被试在听词和观察音韵相关实物词的图片时对不同语音相似条件的注视比例，以探究被试的二语学习经历对一语语音加工的影响。

眼动追踪技术也用于汉字语音的研究，即利用传统的自然阅读的范式，通过操纵汉语形声字的呈现（如省略部分偏旁）探究形声字加工中语音加工的情况。例如，迟慧等（2014）利用眼动技术观察汉语形声字加工中声旁的语音信息的作用，通过观察被试对目标汉字的首次注视、总体注视时间等指标，探究被试在加工汉语形声字时声旁语音信息的影响；Tong et al.（2020）将VWP范式中的图片刺激替换为不同的汉字，通过观察被试对于不同汉字的注视时长等指标，探究汉字识别中形、音、义信息的激活特征。

1.2.4 认知神经科学实验方法

认知神经科学的实验方法能够进一步为语言加工过程提供大脑和神经机制等生理上的证据，在语音加工研究中应用广泛。

事件相关电位（event-related potentials, ERP）研究（或称脑电研究）能够为语音加工过程研究提供加工时间进程等方面的证据。王小璐和程和奔（2021）在关于汉语二语学习者的脑机制研究综述中提到，N170、P200、N400等成分与汉语母语者汉字加工中的声旁特性、语音表征加工等方面有关；而对于汉语为二语的学习者，其汉字的语音加工则不同于母语者，主要表现为N400和LPC（late positive component）在时间进程上相对较晚（Yum et al., 2016）。值得注意的是，该综述也提及，汉字语音加工研究需要区分语音N400和语义N400，以正

确解读汉字语音加工中的脑电研究结果。汉语声调研究中的失匹配负波（mismatch negativity, MMN）（Payne et al., 2020）数据显示，在oddball实验范式中，偏差刺激和标准刺激之间的声学距离大小与MMN的波幅大小存在正相关，与潜伏期负相关 [见王蕴佳等（2021）的研究综述]。

脑成像技术的发展也为语音加工研究提供了更丰富的神经科学证据和研究视角。功能性核磁共振脑成像技术（functional magnetic resonance imaging, fMRI）研究发现，汉字的语音加工涉及左侧额中回（left middle frontal gyrus）、左侧额下回（left inferior frontal gyrus）、左侧顶下回（left inferior parietal gyrus）和双侧枕颞系统 [包括梭状回（fusiform gyrus）和顶中回（middle occipital gyrus）] 等脑区；而拼音文字的加工则主要涉及了腹侧前额系统（ventral prefrontal system）、颞顶背侧系统（dorsal temporoparietal system）和左腹侧枕颞系统（left ventral occipitotemporal system） [详见王小璐、程和奔（2021）关于汉语和拼音文字的fMRI研究综述部分]。另一方面，近年兴起的功能性近红外光谱技术（functional near-infrared spectroscopy, fNIRS）作为一种非侵入式、相对便携的脑成像技术手段，已成为语音加工研究中的一项有效工具。通过检验被试加工不同语音时脑区中的氧合血红蛋白（HbO）、脱氧血红蛋白（HbR）和总含氧量（HbT）等指标，研究者能够观察被试在加工语音时不同脑区的激活程度等情况 [详见杨海波等（2014）针对应用fNIRS技术的语音加工脑机制研究的综述]。

1.3 语音加工的模型和理论

早期的语音加工研究对语音感知的内在机制进行了解释，较为经典的理论和模型包括肌动论（motor theory）和直接现实观（direct realist theory）、队列模型（the cohort model）、踪迹模型（the TRACE model）。现有文献已对这些理论和模型做过详尽解释（如Carroll, 2008; Harley, 2014等），本节仅做简要介绍。

1.3.1 肌动论和直接现实观

言语的肌动论 (Lieberman, 1983, 1985) 认为, 言语感知的对象是语音发音的意图动作 (the intended phonetic gestures), 这些意图动作在大脑中表征为具有语言学特征的发音动作指令 (其中包含了多种语音发音信息, 如元音的圆唇, 辅音的发音部位、发音动作等)。这样的感知模式需要建立语音信号和发音动作之间系统化的联系, 在感知语流时, 听者感知的是语流中发音动作的特征模式。此外, 肌动论的另一个主要的观点是, 大脑中存在独立的语言模块来储存和处理语音感知和加工中发音动作意图的语言学信息。

与肌动论相似, 言语感知的直接现实理论 (Fowler, 1986, 1987, 1996) 或言语感知的生态理论 (ecological theory) (Best, 1994) 关注感知对象的发音, 且更为直接地将声带的发音动作 (the gestures of the vocal tract) 视为言语感知的对象。与肌动论不同的是, 直接现实理论认为, 言语是一般声音感知的一种, 无需独立、特定的感知模块, 听者通过声音信号感知发声的对象和发生的动作。因而, 对于直接现实论来说, 语音的感知无须依赖携带语音信息的心理表征或特定的语音表征作为中介 (Best, 1995)。

肌动论和直接现实观都将语音发音动作作为语音感知的对象或重要成分, 这一观点的必然结果是, 言语感知和言语产出存在互动和联结 (perception-production link), 这对于解释人类在言语交际中出现的协同效应 (convergence 或 imitation) (Lieberman, 1983) 和解释二语学习中二语语音感知和产出的准确性等现象有重要的意义。

1.3.2 队列模型

Marslen-Wilson (Marslen-Wilson, 1975, 1987; Marslen-Wilson & Welsh, 1978) 提出的队列模型将言语感知过程 (尤其是听觉词汇感知过程) 描述为三个阶段——通达阶段 (access stage)、选择阶段 (selection stage)、整合阶段 (integration stage), 认为言语感知过程是

自下而上为主的加工过程。最初的队列模型认为，在通达阶段，词首的语音信息输入首先激活记忆中的语音表征，从而激活了相应词首（如 t）的一组备选词语（如 tree、tea 等），这些备选词语被称为队列（cohort）。在选择阶段，随着语音信息的继续输入，队列中的备选词语逐个被淘汰，最后剩下一个词语，一个词能够区别于所有其他相似词的加工点称为独特性加工点（the uniqueness point）。而到了整合阶段，词汇的语义信息和句法信息才发挥作用，整合到更高单元的言语加工中。例如，在听和感知 trespass（/trespəs/）这个词时，语流中每个音位逐渐被感知，从 /t/ 开始激活大量 /t/ 音开头的词语，到 /tres/ 过程中逐步排除无关词语缩小队列（如 trespass、trest），最后语音输入达到 /tresp/ 时可能达到独特性加工点，这样队列中仅剩目标词 trespass（例自 Harley, 2014）。在自下而上识别言语的基础上，起初的队列模型认为语境影响听力词汇感知的通达和筛选过程，但修正的队列模型（Marslen-Wilson, 1987）认为语境仅在整合阶段起作用。此外，在修正的队列模型中，为了解决声音信号失真和错误感知的问题（如 pocket 听成 bocket），Marslen-Wilson（1987）提出，被排除或淘汰的词语可能会在输入更多语音信息后重新加入队列。

1.3.3 踪迹模型

相比于自下而上加工的队列模型，McClelland & Elman（1986）提出的踪迹模型则认为言语和词汇的识别是一个交互激活的过程。根据踪迹模型，词汇信息由语音特征（feature）、音位（phoneme）和词汇（word）三层众多信息单元（unit）网络构成，被称为“踪迹”（trace）。在感知言语时，踪迹中的信息单元被激活（activate）或抑制（inhibit）。语音的特征信息（如爆破、送气）首先激活特征层中的对应信息单元，随着词汇剩下语音的输入，信息单元的激活继而扩散到音位层和词汇层，最终识别相应的单词。各层级中的激活和抑制是双向的，使得语音感知过程中自下而上的语音特征和自上而下的词汇信息能够共同影响感知加工过程。