心理科学进展 2024, Vol. 32, No. 9, 1488–1501
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2024.01488

·研究前沿(Regular Articles)·

听到“牛黄”能想到“黄牛”吗？——口语识别中的语音位置编码机制*

韩海宾[1] 李兴珊[2,3]

([1]河北师范大学教育学院, 石家庄 050024) ([2]中国科学院心理研究所, 北京 100101)
([3]中国科学院大学心理学系, 北京 100049)

摘 要 在众多语言中, 都存在一系列词汇, 经过语音位置转置后仍能有效成词, 典型如中文中的“牛黄”与“黄牛”。阐明这类可转置词汇在语言理解过程中的编码方式, 是一项至关重要的研究课题。在阅读领域, 学者们已就词汇的位置编码机制展开了讨论, 然而针对口语加工中语音位置编码的认知机制, 至今仍存在序列–灵活编码之争: 早期口语识别理论认为语音位置编码主要以序列编码方式为主, 而近年来的研究则发现, 音位、音节和句子等层面上存在以灵活编码为主的语音位置编码方式。未来研究应深入探索与口语识别中语音编码相关的认知机理、神经机制、语言获得以及人工智能等重要问题, 由于汉字词在形音对应关系和语音加工单元等方面独具特殊性, 后续研究应对汉字词的语音位置编码予以特别关注。

关键词 口语识别, 语音位置编码, 汉字词
分类号 B842.5

1 引言

“研表究明, 汉字的序顺并不定一能影阅响读。” 当你阅读完这句话, 也许都不会发现其字序是错乱的。汉字的位置编码顺序是近些年汉语阅读领域的热门话题。但当你试着读给其他人听的时候, 他们会觉得不知所云。这是为什么？有学者认为, 人眼在阅读过程中会对句子进行“不仔细”的扫视, 大脑会根据语境等自上而下的信息, 并结合自身经验和对后续内容的预测达到阅读理解的目的。阅读过程中, 读者会在其知觉广度范围内, 也就是其注视点周围可获取有用信息的区域, 采取并行加工的方式对文字信息进行处理(Rayner, 1975)。与此不同, 听觉语言理解或者说口语识别过程的主要特点是听觉接收到的言语信号会随着时间线性展开。一个典型的例子能够帮助我们感受到线性输入的口语声音信号: 互联网时代的即时通信工具都具有发送语音的功能, 当我们接收到语音消息的时候, 我们需要点击并收听语音消息, 直到播放完毕后我们才能够通达该语音消息的全部意义。当我们需要快速提取该语音消息的重点时, 我们会使用“转文字”功能。因此, 口语识别区别于阅读理解最大的不同就是其语音信号随时间推移不断输入, 语音信号初始阶段提供的信息不足以让我们识别出口语所传达的意义。我们需要对线性输入的全部声音信号进行声学–语音分析, 并将分析结果映射到心理词典中的词汇表征上(Gaskell & Marslen-Wilson, 1997; Marslen-Wilson, 1990; Marslen-Wilson & Warren, 1994; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Norris, 1994), 进而达到词汇理解的目的。在口语识别中, 很多情况下词汇之间仅仅通过语音的时序位置来区分不同的语义。例如, 当我们识别汉语口语中的“牛黄”与“黄牛”时, 主

收稿日期：2023-10-19
* 河北省社会科学基金项目(HB22JY041)资助。
通信作者：李兴珊, E-mail: lixs@psych.ac.cn

要依赖的就是这两个音节在时间上出现的不同顺序。因此, 语音的时序位置编码在口语识别过程中扮演着至关重要的角色。

针对该问题, 以往大部分研究者认为口语识别采用线性序列编码的加工方式, 词汇的语音表征也是基于固定的语音序列(Chen & Mirman, 2012; Gaskell & Marslen-Wilson, 1997; Marslen-Wilson, 1973, 1985; Norris & McQueen, 2008)。例如, 声音特征、音位、音节等在词汇语音表征中的先后位置。近 5 年, 有研究者注意到在拼音语言中, 口语词汇识别过程的语音编码方式似乎更为灵活, 可以脱离其固定的位置单独进行编码(Dufour et al., 2022; Dufour & Grainger, 2019, 2022; Gregg et al., 2019; Mirault et al., 2018; Toscano et al., 2013)。例如, 在拼音文字中, 语音位置转置后的词汇可以互相激活(Toscano et al., 2013), 听到“sub”之后, “bus”也会得到激活。那么, 对于汉语口语识别过程来讲, 按照语音灵活编码的研究结果, 我们在听到“牛黄”后, 其可转置词“黄牛”会得到激活吗?

针对这些现象, 研究者们关注的焦点问题是口语识别过程中语音时序位置编码的内在机制。作为心理语言学领域的重要议题, 厘清口语加工的语音编码机制对于揭示人类认知过程具有重要的理论意义。本文将重点综述近年来针对口语加工过程中语音位置编码的研究进展, 对比分析两种不同的语音编码方式。首先, 全面回顾口语识别和加工的序列位置编码理论和相关研究; 其次, 介绍在音位、音节和句子等不同水平上, 口语识别过程中语音位置的灵活编码理论和最新研究进展; 最后, 展望未来关于汉字词语音位置编码在口语识别中的特殊性, 并探讨将来可能的研究方向。

2　口语加工的序列位置编码理论: 基于位置的语音编码方式

序列位置编码模型作为口语加工过程的重要理论, 已经得到大量研究的支持, 能够解释许多口语加工中的现象(Allopenna et al., 1998; Luce et al., 2000; Marslen-Wilson & Tyler, 1987; Marslen-Wilson & Zwitserlood, 1989; McClelland & Elman, 1986; Norris, 1994)。该类理论模型基于一个客观事实: 词汇的语音是一个有序的信息序列, 即单词开头的声音信号比后面的声音信号先到达听者的耳朵。鉴于此, 学者们认为听者首先对言语的声音信号进行声学−语音分析, 随后将输出的结果映射到心理词典中词汇的语音表征上, 达到词汇通达的目的(Gaskell & Marslen-Wilson, 1997; Marslen-Wilson, 1990; Marslen-Wilson & Warren, 1994; Marslen-Wilson & Welsh, 1978)。研究者们构建了诸多模型来模拟口语识别过程, 虽然各类模型在细节上有差别, 但他们都共享一个假设: 个体按照它们在语音输入中的先后位置进行编码, 最终成功地将其映射到这个有序的信息序列表征上, 从而实现对词汇的识别。

作为时序性模型的代表, Marslen-Wilson 和 Tyler (1987)构建的 Cohort 模型, 也称词汇识别群集模型, 常被用来解释口语词汇的识别过程(见图 1)。早期该模型认为词汇的识别完全依赖于自下而上输入的听觉信息, 词汇词首的声音特征或音位会激活一组词首语音与之精确匹配的候选词汇(cohorts)。这种激活遵循“全或无”的原则, 词首的语音信息将直接决定候选词汇组的大小, 词首语音可以匹配的将继续参与后续词汇的识别过程, 不匹配的则不参与, 且会被移除出候选词汇组。随着语音信息自下而上的不断输入, 语音不一致的词汇逐步被排除, 最后识别出所对应的词条, 进而识别出对应的词汇。这种序列模型或者说线性加工模型得到了大量研究的支持(Marslen-Wilson & Zwitserlood, 1989; Marslen-Wilson et al., 1996)。Marslen-Wilson (1973)最早使用影子跟读任务(speech shadowing task)去测量句子中每个单词的响应延迟, 该任务要求被试听取录音并尽可能迅速地大声重复(跟读)所听到的内容, 结果发现跟读者往往在一个词还没有完全播放完之前就开始重复它, 这表明跟读者心理词典中的词在听到整个词之前就已经被该单词的起始音位激活了。诸如此类的发现也促使 Marslen-Wilson 和 Tyler 于 1987 年提出了 Cohort 模型。Marslen-Wilson 和 Tyler (1987)的研究中指出虽然“**c**ategory”与“**g**ategory”只有词首音位不同, 但也不会互相激活。然而, 后续研究发现, 即使是词首音位不匹配的词汇, 只要在共同的位置有匹配的语音信息, 也会被激活(Allopenna et al., 1998; Chen & Mirman, 2012; Gaskell & Marslen-Wilson, 1997; Norris & McQueen, 2008; Dufour & Peereman, 2003; Marslen-Wilson et al., 1996; Zwitserlood, 1989)。例如, Allopenna 等人(1998)使用视觉情境范式考察了口

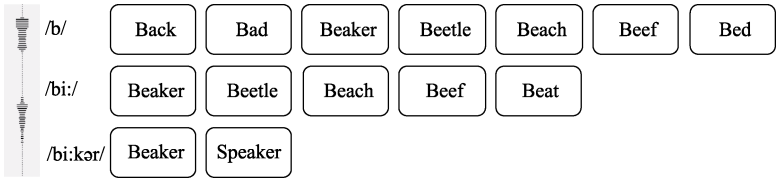

图 1 Cohort 模型识别听觉词汇“Beaker/bi:kər/烧杯”的心理过程。第一行为听到音位/b/后激活的起始音相同的词汇组; 第二行为听到音节/bi:/后激活的一系列词汇, 此时与/bi:/不匹配的词汇已经被移除; 最后当整词语音结束后, 除 Beaker 以外的词汇全部被移除; 但后期修正的模型发现, 韵脚所在位置的音节也会激活韵脚相同的词汇。

语识别过程。该实验给被试听觉呈现词汇“Beaker” (烧杯), 屏幕上给被试呈现 4 个物体, 分别对应目标词“Beaker”, 词首音匹配词“Beetle” (甲虫), 韵律音匹配词“Speaker” (扬声器), 无关词“Carriage” (婴儿车)。结果发现被试在听到 Beaker 的时候, 眼睛先注视了 Beaker 和词首音匹配的 Beetle, 但随着语音信号的输入, 也注视到了韵律匹配的 Speaker, 该研究与 Cohort 模型提出的“全”或“无”的假设是不一致的。实际上, Marslen-Wilson (1993)后期对其模型进行了修正, 主张不匹配的词汇不会被完全排除, 而是被延迟激活, 听觉词汇的识别不是“全”或“无”法则, 最后识别出的词汇也只是激活程度高于其他词汇。但其依然主张自上而下的信息只能在整合阶段起作用, 不参与最初的词汇识别过程。

相比 Cohort 模型, Allopenna 等研究者的结果更多地支持了持交互观点的持续匹配模型(continuous mapping models), 例如 TRACE 模型(McClelland & Elman, 1986), NAM 模型(Luce et al., 2000) 和 Shortlist 模型(Norris, 1994) 等等。McClelland 和 Elman 在 1986 年提出了最具代表性的词汇识别 TRACE 模型, 该模型是交互激活模型的代表, 主要有两点主张: (1)听觉词汇的识别依赖于听觉输入的语音信号和心理词典中词汇语音表征的整体匹配程度, 而不仅仅是依靠词首语音信息; (2)自上而下的信息从最初阶段就参与了词汇识别过程, 而不是到整合阶段再起作用。该模型还强调, 整个模型映射网络内的声音特征水平、音位水平和词水平之间的联系是交互的, 可以向上或者向下传递。该模型和 Cohort 早期模型不同, 但同样也得到了很多研究的支持(Connine et al., 1993; Connolly & Phillips, 1994; Dahan & Magnuson, 2006; Sereno et al., 2003; Van Petten et al., 1999)。例如, Connine 等人(1993)采用跨通道语义启动范式发现原词汇启动效应最大, 与原词汇只有一两个特征不匹配的非词启动效应居中, 而有 6 个特征不匹配的非词没有启动效应, 研究表明, 听觉刺激与心理词典表征的相似性程度决定了词汇的识别。但在 TRACE 模型中, 每次新的特征输入时都需要复制整个模型映射网络。为了解决这一时间成本问题, Norris (1994)提出了一种听觉词汇识别的混合模型——Shortlist 模型。该模型包括两个关键阶段: 第一阶段采用独特的自下而上加工方式, 通过详尽的串行词汇搜索, 激活与输入匹配的单词候选短列表; 第二阶段的交互激活方式类似于 TRACE 模型的词汇层。与 TRACE 模型相比, Shortlist 模型在词汇通达的第一阶段仅允许自下而上的影响。但相较于 Cohort 模型, 它又避免了严格的时间顺序约束, 使得即使在起始部分不匹配但在其他部分匹配的候选词也可以被激活。而 NAM 模型(Neighbourhood Activation Model, Luce et al., 2000)更关注目标词在任何语音位置上的竞争者——“邻居”。语音邻居指的是一组在语音上相似的词汇列表, 例如“hide” (躲藏)和“tidy” (整洁)都是“tide” (潮汐)的邻居。该模型认为输入的声音信息与这些“邻居”的匹配也会激活相关词汇, 例如, “us” (我们)可以作为“bus” (大巴)语音/字形相似的“邻居”而被激活。

通过梳理这些文献, 我们发现早期理论模型存在差异(具体列表见表 1), 且都有大量实验证据的支持。我们也发现虽然它们存在差异, 但这些理论也都呈现一个共同的特点: 在口语加工过程中, 词汇由序列编码的语音片段(例如, 声音的细节特征, 音位, 音节等等)构成。这些词汇的激活和识别源于输入的声音信号与这些序列语音片段的线性匹配程度。具体而言, 词汇的语音信息都被放进固定的“槽(slot)”中。以“bus/bʌs/”这个词为例, 假设存在三个分别储存不同音位的“槽”, 其

表 1　早期词汇识别的代表性序列模型对比

代表模型	识别方式	识别过程	识别要素	自上而下信息起作用阶段
Cohort 模型	自下而上	严格序列性	词首信息	后期整合阶段
TRACE 模型	自下而上与自上而下	交互激活	词汇与心理词典匹配的整体效应	自始至终
Shortlist 模型	自下而上与自上而下	交互激活	词汇与心理词典匹配的整体效应	单词候选列表阶段之后的选择阶段
NAM 模型	自下而上与自上而下	交互激活	词汇的“邻居”的整体相似性	词汇“邻居”激活后的词汇决策阶段

先后位置是固定的。当/b/和/ʌ/这两个槽匹配时, “bus”、“butter”等词汇都可能激活; 同样地, 当/ʌ/和/s/这两个槽匹配时, “us”可能会激活。这些理论被称作基于槽框架的理论(slot-based theory), 语音和位置信息的编码是固定的, 词汇的激活依赖于每一个槽的语音和位置的匹配程度(Gaskell & Marslen-Wilson, 1997; Luce et al., 2000; Luce & Pisoni, 1998; McClelland & Elman, 1986; Norris & McQueen, 2008)。

然而, 在阅读或者视觉单词的识别领域, 早有研究者发现词汇字母的编码顺序可能采用一种粗略编码(coarse-encoding)的方式(Chambers, 1979; Grainger & Whitney, 2004; Perea & Lupker, 2003, 2004)。例如, 真词与任意交换两个字母而形成的假词“ga**rd**en” (花园)和“ga**dr**en”, 被试做出词汇判断比替换该真词中两个字母而形成的假词“ga**ts**en”更困难(Chambers, 1979; Frankish & Turner, 2007; O'Connor & Forster, 1981)。Perea 和 Lupker (2003)也发现字母位置转置后的假词“ju**gd**e”可以启动其对应的原始词“ju**dg**e” (评判), 但替换原始词中两个相同位置的字母“ju**lp**e”则不能激活。他们后续研究也发现该现象也适用于非相邻字母的转置(例如, ca**ni**so/casi**n**o, Perea & Lupker, 2004), 研究表明, 字母位置放置错误的词汇仍然可能激活正确的目标词。字母位置信息的编码具有一定程度的位置不确定性(Gomez et al., 2008)以及灵活性(Davis, 2010; Grainger & van Heuven, 2004; Whitney, 2001)。但是, 这并不意味着读者可以完全忽略字母的顺序, 而且这些效应存在一定的限制(Guerrara & Forster, 2008; Hannagan et al., 2011)。在句子水平, 也有研究者提出, 阅读者在理解句子过程中对刚刚阅读过的单词保留一种不确定性, 这种不确定性在句子阅读过程中会一直保持(Levy et al., 2009)。

早期关于口语识别模型的研究多认为词汇的语音位置编码是固定的、线性的序列编码, 词汇的识别依赖于输入的语音信号序列与心理词典中语音表征的精确匹配程度。然而, 阅读领域中关于“不确定性”的研究结果给了我们启示。不论是词汇还是句子, 在字母或者词的编码顺序方面可能采用粗略编码的方式, 其位置信息的编码存在一定程度的不确定性。尽管口语加工过程和书面语言在信号输入的通道、性质和时间进程上都不太相同, 但这些研究无疑为我们提供了一种在口语加工中重新思考语音时序位置编码可能性。

3　口语加工的灵活位置编码理论: 独立于位置的语音编码方式

不同于采用序列位置编码的口语识别模型, 近期有研究者开始探索口语识别中相对灵活的语音编码方式(Dufour & Frauenfelder, 2010; Dufour et al., 2022; Dufour & Grainger, 2019, 2022; Gregg et al., 2019; Liu et al., 2020; Mirault et al., 2018; Toscano et al., 2013)。结合新构建的理论模型, 例如, TISK 模型(Hannagan et al., 2013; You & Magnuson, 2018), 这种独立于位置进行编码的语音编码方式逐渐崭露头角。在这一部分, 我们将梳理近期关于口语识别中在音位、音节以及句子等水平上探索语音位置灵活编码的研究, 并对相应的理论模型加以介绍。

3.1　音位水平的语音位置编码

最早关注该问题的研究出现在 2013 年, Toscano 等人注意到音位(Phoneme)在词汇的不同位置上具有非常细微的声学差异, 听者对这些差异非常敏感(McMurray et al., 2002)。他们认为这些细微差异的存在使得每个位置读音会有细微不同, 所以被试无需对声音信号进行位置编码即可进行词汇识别。因此, Toscano 等人(2013)使用视觉情境范式(visual world paradigm)考察了口语加工中的音位位置编码机制。视觉情境范式由

Cooper (1974)首次使用, 该范式的特点是听觉与视觉能以更自然的形式共同呈现, 其精确的时间锁时特性能够帮助研究者们更好地研究口语加工过程(Andersson et al., 2011; Cooper, 1974; Huettig & Altmann, 2005; Han & Li, 2020; 韩海宾 等, 2019)。该范式同时向被试在屏幕上呈现视觉刺激和听觉上呈现声音信号, 被试“听声音”的同时完成“看屏幕”或者点击屏幕目标位置的任务, 最后根据眼动轨迹推测口语加工的内在机理。在 Toscano 等人的实验中, 听觉词汇为“sub” (次级), 屏幕上呈现目标词“sub” (次级)、音位转置词“bus” (大巴)、词首音一致词“sun” (太阳)以及无关词“well” (好)。结果发现, 被试对目标词的注视概率最高, 其次为词首音一致条件(cohort)。最重要的是, 作者发现了音位转置效应(transposed-phoneme effect), 即被试对音位转置词的注视概率显著高于无关词。不仅如此, 作者还进一步将实验材料细化到了声音特征水平, 例如, 作者发现虽然“tack” (钉)与“cat” (猫)并不互为音位转置词, 但是其音位转置后因为有相似发音特征的/t/和/k/, 也出现了音位转置效应。Toscano 等人认为在口语词汇的加工中, 词汇的语音并不是以固定的“槽”表征的, 音位位置颠倒的词汇也会得到编码和激活。他们同样主张词汇的语音表征并没有固定的位置编码, 其对结果的解释是编码基于一系列精确和细致的语音特征线索, 并没有注意到语音位置编码的特殊性。

为了验证该效应, 有研究改进并重复了 Toscano 等人的实验(Dufour & Grainger, 2019, 2020; Gregg et al., 2019)。例如, Gregg 等人(2019)认为 Toscano 实验中出现音位转置效应的原因可能是“bus”和“sub”中元音/ʌ/的重叠, 因此, 除了重复音位转置效应, 作者还增加了无元音位置重叠的条件“leaf” (叶子)和“flea” (跳蚤), 结果发现无元音重叠条件并没有发现转置效应。作者没有否认语音位置编码存在一定的灵活性, 但作者认为至少元音位置非常重要, 口语词汇识别模型应该探索更加专业的编码系统, 应兼具位置的匹配与灵活性。

在视觉情境范式的研究中, 存在一个明显问题: 目标词和转置词在屏幕上同时出现, 这很可能会导致观察到的转置效应与视觉词的转置有关。Dufour 和 Grainger (2019)改用启动范式考察了口语词汇识别中的音位位置编码, 并设置了更多的实验条件。实验中的目标词读音为/tyb/, 重复条件中的启动词和目标词一致, 音位转置条件中的启动词读音为目标词的转置词/byt/, 控制条件则音位无关/mul/。为了排除元音位置的作用, 作者还设置了仅有元音位置重叠的条件(/jyp/和/dyn/)、两个重叠音位的条件(/byl/与/tyb/), 结果发现了音位转置效应, 即音位转置条件的反应时要显著短于控制条件(/byt/可以启动/tyb/), 其他条件并不显著。

这些研究表明, 在一定程度上, 相同音位构成的词汇即使音位位置不同也可以相互激活, 口语识别过程中的音位可能独立于其位置信息进行编码, 语音的编码采用的是一种粗略的、更为灵活的编码方式。

3.2 音节水平的语音位置编码

严格来说, 还尚未有研究考察在口语识别中音节水平上的语音位置编码, 也就是将一个听觉词汇中的音节(syllable)位置颠倒后是否会出现音节转置效应(transposed-syllable effect)。但有研究者考察和对比同一音节内音位转置效应与跨音节间音位转置效应(Dufour & Grainger, 2022; Dufour et al., 2021)的差异。虽然这些研究实际考察的依旧是音位编码, 由于这些研究的重点主要是强调音节内以及音节间语音编码的差异, 本综述将其划分入音节水平。

Dufour 等人(2021)采用启动范式, 通过真假词判断任务, 考察同音节内的音位转置效应以及跨音节间的音位转置效应的差异, 以此来揭示口语词汇识别中在音节水平上音位的位置编码机制。实验中构建了: 转置的音位同属一个音节的音节内转置条件(例如, /bis.**tɔk**/和/bis.**kɔt**/, “.”为音节分隔符)和转置的音位隶属两个音节的跨音节转置条件(例如, /ʃo.**lo**.**ka**/和/ʃo.**ko**.**la**/)。除此之外, 为了平衡音节前后位置效应, 作者还将音节位置设置为一个变量, 前音节位置条件(例如, /**sɪb**.kɔt/和/**bis**.kɔt/)和后音节位置条件(例如, /bis.**tɔk**/和/bis.**kɔt**/)。结果发现, 转置条件下做出假词的判断要比控制条件长, 出现了音位转置效应。作者还发现音节内转置和音节间转置的效应并不显著, 音节前后位置的效应也不显著。研究表明, 语音的位置编码在音节水平上也具有灵活性, 这种灵活性编码不局限于音节之内的转置, 还可以跨音节发生。作者的研究进一步说明音位编码是独立

于位置进行的, 贯穿整个听觉词汇刺激的输入过程。Dufour 和 Grainger (2022)使用短时启动范式重复了该实验, 并在此基础上增加了对两个音位在词汇中距离远近的考察。具体来说, 作者除了设置近音位距离条件(例如, /bi**ks**ɔt/和/bi**sk**ɔt/)和远音位距离条件(例如, /ʃo**ɬ**o**k**a/-/ʃo**k**o**ɬ**a/)。结果作者在近音位距离条件下发现了比无关条件更长的反应时, 但远音位距离则没有发现该效应, 作者认为可能是由于远近距离的对比使得近音位条件下的效应更强导致的。

总的来说, 在音节水平, 研究者发现不管是同音节内的音位, 还是跨音节间的音位, 都具有转置效应。音位的编码在音节水平上同样是灵活的, 是独立于输入位置进行的。然而, 尚无研究者直接考察音节的语音位置编码。根据目前有关音位和音节编码的研究来看, 实际上都属于在音位层面进行的考察, 其结果也都支持了一种独立于输入时序位置的编码机制。正如下文对 TISK 模型的介绍中提到的, 我们会对输入的语音序列进行编码, 形成一系列与位置无关的音位或者二音位矩阵, 然后在词汇水平层面或词汇决策阶段再识别出语音重叠程度较多的词汇。

在音位与音节位置编码的神经机制研究方面, Yee 等人(2008)发现额下回(Inferior Frontal Gyrus, IFG)的病变使得词首起始音位的竞争效应变得非常小, 而颞上回(Superior Temporal Gyrus, STG)的病变则导致该竞争效应比对照组更大。同时, 其他研究者发现左侧缘上回(Supramarginal Gyrus, SMG)和左侧额下回参与了词汇的语音竞争(Prabhakaran et al., 2006; Righi et al., 2010)。Yi 等人(2019)以及 Scott (2019)都强调了左侧颞上回可能参与了语音特征空间和时间上在大脑中的整合过程。我们猜测, 语音位置编码可能与左侧颞上回相关。然而, 目前对于语音位置编码的神经机制的研究还相对较少, 为了更深入地揭示音位和音节位置编码机制, 需要更多的神经科学研究来支持对语音编码机制的深入理解。

3.3　句子水平的语音位置编码

很多学者考察了阅读过程中句子水平上词汇位置的编码机制, 部分研究者认为句子中词汇顺序以及位置的编码是序列进行的(Reichle et al., 1998), 也有一些研究者认为阅读者在句子理解的过程中会一直对先前阅读的单词保持一种不确定性(Levy et al., 2009)。不管是拼音语言(Mirault et al., 2018), 还是汉语这种表意语言(Liu et al., 2020; Liu et al., 2021; Liu et al., 2022), 编码是相对灵活的。但这些研究都是针对阅读开展的研究。

我们加工口语词汇的速度相对加工句子来讲, 速度非常快, 对音位的处理速度也更快。因此, 作为词汇位置底层元素的音位在自下而上输入过程中很容易受到嘈杂的听觉输入的“噪音(noise)”的影响(Dufour et al., 2022)。在音位水平发现的灵活编码可能是由于这种“噪音”造成的模糊位置匹配的结果, 这也可能是音位转置效应出现的原因之一。因此, 在句子水平, 词汇与词汇之间有间隔, 相比于音位水平, 受“噪音”的影响会大大降低, 考察口语句子理解中的词汇转置效应(transposed-word effect)更能为语音的灵活编码理论提供证据。但很少有研究针对句子水平的语音编码进行探讨, Dufour 等人(2022)考察了口语句子理解过程中的语音位置编码现象。该实验采用语法决策任务(grammatical decision task), 被试会听到由五个词汇构成的句子, 并且需要判断该句子的语法是否正确。实验设置了语法正确的句子“The black **dog was** big” (这个黑色的狗很大), 可以通过交换单词变成语法正确句子的词汇转置条件“The black **was dog** big”, 以及不能通过交换单词变成语法正确句子的非转置条件“The black was dog slowly”。结果发现, 与非转置条件相比, 词汇转置条件做出不符合语法判断的时间要显著更长, 且错误率更高。作者首次证明了句子中的词汇转置效应不仅在阅读中存在, 也同样存在于口语加工过程中, 口语句子加工中词汇转置效应的出现源自于语法和语义的双重限制。

句子加工的理论通常假设输入我们语言处理机制的句子是一个准确无误的词汇序列(Hale, 2001; Jurafsky, 1996; Levy, 2008), 实际上字词缺失、听力受损、环境等“噪音(noise)”普遍存在于语言的日常使用中。Gibson 等人(2013)提出的言语感知的噪声通道理论(Noisy-channel model)就解释了我们如何从“噪音”的影响中理解语言。然而, 大部分支持这一理论观点的研究都来源于阅读领域。但读者或者听者试图从嘈杂的语言输入中理解句子的思想很好地解释了词汇位置效应的出现。语音位置的灵活编码理论无疑为噪声通道理论是否能够处理听觉通道的句子提供了更有意

义的证据。

因此，在听觉句子的加工中，类似于读者在阅读理解过程中对刚刚阅读过的单词保留一种不确定性一样(Levy et al., 2009)，大脑对语音的加工同样采用一种独立于具体词汇语音位置的编码机制。神经机制层面的研究也支持了类似的结果，发现大脑可以快速解码语音特征，而语音位置顺序信息是独立存储的。例如, Gwilliams 等人(2022)通过脑磁图记录了 21 名参与者在听短篇故事时近两个小时的大脑活动，探讨了人类大脑如何处理和理解语音这种连续变化的声学信号。研究发现，人类大脑能够持续并行地处理和编码最近听到的三个语音单元。这意味着大脑不仅在声音出现时才对信息进行加工，即使这些声音已经从感觉输入中消失，大脑仍然保持它们的信息。这些信息中语音的具体特征和自从声音开始以来所经过的时间顺序信息是分开进行加工的。这些研究为我们独立于时序位置进行的语音编码机制提供了支持。

3.4 语音位置编码的影响因素

音位转置效应并非在所有情况下都有发现，研究者从不同的角度通过转置效应考察了语音位置编码机制，发现音位距离、词频、感觉通道等特征都会影响语音编码过程。

首先，词汇频率会影响口语识别过程中音位转置效应的出现(Dufour & Grainger, 2020)。Dufourh 和 Grainger (2020)采用短时语音启动范式，通过操作启动词和目标词的词汇频率来考察词汇频率信息在音位转置启动效应中的作用。结果发现当目标词比启动词的词汇频率高时，会出现音位转置效应；当目标词的词频低于启动词时则不会。研究表明，词汇频率信息也会自上而下的影响口语识别中的语音编码过程。作者认为，在启动词加工过程中，其词汇表征被部分激活，影响了目标词的加工。

其次，在前面探讨音节水平的语音编码时，我们提到了 Dufour 等人(2021)对同音节内音位的转置效应以及跨音节间的音位转置效应的考察，其实验中采用的实验材料(例如，/bis.**tɔk**/ 和 /bis.**kɔt**/，/ʃo.**l**o.**k**a/和/ʃo.**k**o.**l**a/)的转置都间隔一个音位，结果发现了音位转置效应。但 Dufour 和 Grainger (2022)采用短时启动范式重复了该实验，并设置了近音位距离条件(例如，/bi**ks**ɔt/和/bi**sk**ɔt/)和远音位距离条件(例如，/ʃo**l**o**k**a/和/ʃo**k**o**l**a/)，结果显示远音位距离没有发现音位转置效应，只在近音位距离条件下发现了比无关条件更长的反应。虽然目前结果并不一致，但在一定程度上表明转置音位在词汇中的距离可能会影响口语加工过程中的语音编码过程。

再次，有研究者发现音位转置效应可能受到启动词与目标词呈现的感觉通道的影响。例如，Dufour 等人(2023)对比了不同感觉通道的呈现方式对音位转置效应的影响。在同一种感觉通道的实验中，启动词和目标词都是以听觉呈现；在跨感觉通道的实验中，启动词采用视觉词汇呈现，目标词采用听觉通道呈现。结果发现在同种感觉通道呈现的实验中具有促进效应，而在跨通道的实验条件下会出现抑制性的启动效应。

最后，我们想强调的是目前拼音语言中语音位置编码研究存在的问题。通过梳理和阅读近期关于转置效应以及口语识别中语音位置编码的文献，我们发现，无论是 Toscano 实验中的“sub”和“bus”，“TACK”和“CAT”，抑或是 Gregg 实验中的“tab”和“bat”，以及 Dufour 等人实验中使用的“LOBE”和“BOL”，均存在一个拼音语言中普遍的问题：形音不分离。其音位等语音信息对应着固定的字母或字母组合(/f/-ph，/t/-t)，具有正字法上的表征，这种形音之间较强的连接关系可能会对口语词汇的加工产生影响。尤其是 Gregg 等人(2019)以及 Toscano 等人(2013)采用的方法是视觉情境范式。该范式同时向被试呈现视觉和听觉刺激，并要求被试“听声音看屏幕”，最后根据口语加工过程中的眼动轨迹推测口语加工的内在机理。这些使用该范式考察音位转置效应的研究，屏幕中的音位转置词“bat”和目标词“tab”会预先呈现给被试一段时间，之后被试才会听到目标词“tab”。被试在加工“bat”的过程中，其字形语音等都会得到很强的激活，目标词与转置词在视觉上的同时呈现更加强了这种效应。因此，不仅仅是拼音文字的形音不分离特性，以往研究可能也存在方法上的问题。也因此，有研究者采用短时语音启动范式等启动范式来考察口语识别过程(Dufour et al., 2022; Dufour & Grainger, 2019)。

3.5 音位灵活编码的新模型——TISK 模型

TISK 模型(Time Invariant String Kernel Model of Spoken Word Recognition, Hannagan et al., 2013;

You & Magnuson, 2018)能够较好地解释上述研究中发现的现象, 即口语词汇加工中独立于位置进行的语音编码方式。同时, TISK 模型也是现有的唯一能解释音位转置效应的模型。同 TRACE 模型一样(McClelland & Elman, 1986), TISK 模型也是一种交互激活模型, 但其所持的基本观点不同。TISK 模型的提出来源于机器学习领域中一种非常成功的计算技术——字符串核函数(Hofmann et al., 2008)。这种字符串核函数可以独立于字符串中符号的位置来比较符号序列, 字符串核函数可以表示为符号组合的高维空间中的点。例如, 英文词汇“TIME”可以看成一个字符串序列, 作为一个向量, 其中每个分量是两个符号的组合, 即“TI”、“TM”、“TE”、“IM”、“IE”、“ME”。基于此, TISK 模型假设词汇的语音是用一组位置无关的音位单元(phoneme, 例如“T”、“I”、“M”、“E”)和一组代表连续和不连续音位序列的开放二音位单元(open-diphone units, 例如“TI”、“TM”、“TE”、“IM”、“IE”、“ME”)构成的。

在这样的框架下, 我们不会编码音位的精确位置顺序, 位置无关的音位单元和开放二音位单元表征都可以为音位转置效应做出贡献。图 2 显示了 TISK 模型下口语识别过程的示例。该模型共分为 4 个水平: 输入水平、音位水平、单音位与开放二音位水平, 以及词汇水平, 其中音位水平是按照时序位置编码的, 但是单音位与开放二音位水平、词汇水平是与时序位置无关的。整个过程从言语声音信号输入开始, 需要经历一组音位单元激活构成声音信号的音位, 其次到一组与输入时序位置无关的单音位与开放二音位组成的矩阵, 第三个水平为词汇水平, 与之可匹配的词汇会得到激活。

TISK 模型得到了近期关于灵活编码的研究的支持(Dufour & Grainger, 2019, 2022; Gregg et al., 2019; Toscano et al., 2013)。举例来说, 输入的刺激词汇/ba**ks**ɛt/将激活一组开放二音位单元, 比如 b-a、b-k、b-s、a-k、a-s、k-s、k-t、t-k 等等, 其中许多与其转置单词 /ba**sk**ɛt/ 兼容。因此, 给定的音位序列与对应单词之间的重叠程度, 无论是位置无关的音位(相同的音位构成)还是开放二音位单元(匹配程度高于控制条件), 这都决定了音位转置效应的发生。Dufour 和 Grainger (2022)的研究很好地支持了该效应, 在其研究中, 作者发现比起替换掉音位的非词/bi**pf**ɔt/, 音位转置词/bi**ks**ɔt/会促进目标词/bi**sk**ɔt/的加工, 就是因为/bi**pf**ɔt/这样的非词会激活许多与/bi**sk**ɔt/不兼容的语音单元。除此之外, TISK 模型并没有完全抛弃语音输入的时序问题, 在音位水平是具有时间序列性的, 这表现在音位距离参数上。例如, Dufour 和 Grainger (2022)发现音位的时序距离较远时(/ʃo**l**o**k**a/-/ʃo**k**o**l**a/), 没有出现转置效应。所以 TISK 模型在参数的设置上, 语音输入生成的开放二音位单元表示的集合是由距离参数控制的。在这个例子中, 音位输入之后其组合而成的矩阵里, 开放二音位单元/lk/和/kl/可能超过了其时序距离参数, 因此没有发现转置效应, 而相邻的音位组成的开放二音位单元则可以得到激活。

4　汉语口语识别中语音位置编码的特殊性

综合以上内容可见, 目前关于口语词汇识别模型的证据主要来源于对拼音文字系统的研究。

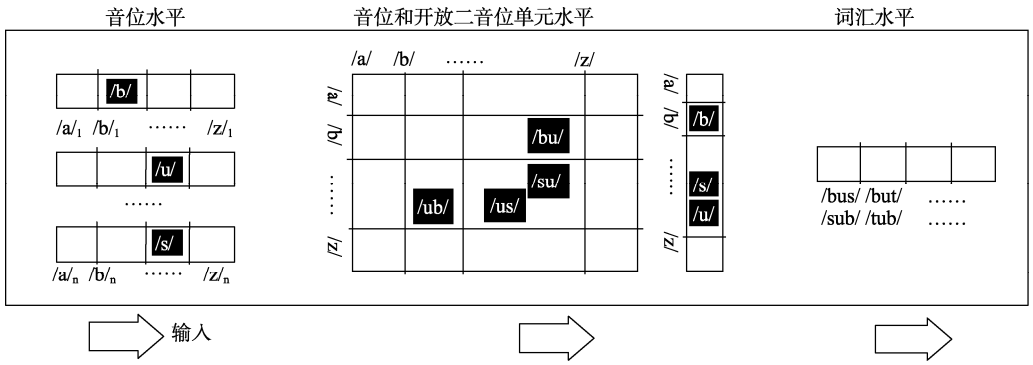

图 2　TISK 模型口语识别过程示例。首先, 从言语声音信号输入开始, 需要经历一组音位单元激活构成声音信号的音位; 其次到一组开放二音位单元水平, 该水平是一组与输入位置无关的开放二音位; 最后为词汇水平, 与之可匹配的词汇得到激活(Hannagan et al., 2013)。

然而，每种语言都具有独特的特点，可能在加工机制上存在差异。为了验证理论的普适性，通常需要对多种文字系统进行跨语言研究，来真正深入探讨人类听觉语言加工的本质。作为一种表意文字系统，汉语有其独特的特点，这部分我们主要强调汉语口语识别中的三点特殊之处。

首先，与拼音语言不同，汉字词具有形音分离(spelling-sound dissociation)的特性。汉字的发音和字形之间的对应关系并不总是规律明确，字形和发音之间的关系不容易被推测或理解，透明度与拼音文字相比较低。前面提到，拼音文字系统最大的问题就是形音不分离，因此，形音之间较强的连接关系可能会对口语词汇的加工产生影响。尤其是当使用视觉情境范式的过程中，字形语音等都会得到很强的激活，目标词与转置词的同时呈现也增强了这种效应。即使后期有研究者采用短时语音启动范式来考察，但依然无法排除形音联结的问题。汉字中我们可以找到读音互为转置词，但字形完全不同的词对(例如，“岩石”和“食盐”，“追兵”和“冰锥”)。汉字词语音、字形联结规则的特殊性可以使我们排除字形的影响去考察语音的编码，为揭示口语词汇识别中语音的位置编码提供更为精确、细致的材料。

其次，汉语词汇的语音加工单元较为特殊。在拼音语言中，有研究者认为词汇的表征由音位组成，音位首先被提取出来(Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Norris, 1994)。也有研究主张言语信号直接和词汇表征相联系，不存在音位中介(Gaskell & Marslen-Wilson, 1997; Lahiri & Marslen-Wilson, 1991; Marslen-Wilson & Warren, 1994)。但目前的研究焦点都在围绕音位水平的编码情况进行探讨，也确实证明了音位表征的心理现实性以及音位编码的灵活性。在上述总结“音节水平编码”的过程中，我们发现目前拼音文字还未真正探讨音节的转置效应和编码情况，只是在探讨音节内或跨音节的音位编码。较之拼音文字，汉语比较独特，其语音加工单位可能以音节为单元(Chen et al., 2002; You et al., 2012)，每个汉字对应一个音节。但也有研究发现汉语母语者可以区分不同的音位，例如，/ma/和/fa/的差异(黄伯荣，廖序东, 2011)。Qu 等人(2012)关于言语产生的研究也为汉语言语产生中的音位加工提供了证据。汉语的语音加工单位可能为“具有音位表征的音节” (O’Seaghdha et al., 2010)。因此，汉字词不仅可以探讨口语识别中的音位位置编码，以及这种音位编码是否独立于其位置而进行(例如，“/an4/”是否会激活“那/na4/”)。其次，汉语词汇以两字复合词偏多，也可以更好地考察口语识别中的音节位置编码机制(例如，“追兵/zhui1 bing1/”和“冰锥/bing1 zhui1/”)。

最后，汉语口语识别过程中音节–语素–语义之间关系的特殊性也是与拼音文字的重要区别。对于汉语来说，音节、语素与汉字通常是对应的，但一个音节可能对应多个语素与汉字，包含了丰富的语义信息。在汉语口语识别过程中，首音节往往已经可以提供丰富的语义信息，例如，听到/bingzhui/中的/bing/已经能够激活多个语素和语义。这和拼音文字不同，拼音文字听到/tæk/中的/tæ/后，是几乎无法提供任何语义信息的。而以往研究发现，语义信息可以促进词汇的识别过程(彭聃龄 等, 1999)。因此，汉语口语识别过程中词首音节提供的语义信息可能在早期就对后续语音的识别产生了自上而下的影响。

目前，拼音文字的研究中，针对口语识别灵活编码的实证研究基本都支持了 TISK 模型，该模型对这些拼音文字系统中音位转置效应的结果也具有较好的解释力度。目前还尚未有研究考察汉语口语识别中独立于位置进行编码的语音加工机制。汉语的语音加工单元为音节，例如/bing1zhui1/中的/bing1/则为一个加工单元，按照该模型框架，我们不会编码汉字音节单元的精确位置，作为位置无关的音节/bing1/、/zhui1/与开放的二音节单元/bing1 zhui1/、/zhui1 bing1/都可以对/zhui1bing1/的激活做出预测。我们猜测，这在汉语的成语中可能表现更为明显。诸如“事半功倍”和“事倍功半”这样的词汇在口语识别中容易引发混淆，可能就是因为/shi4 ban4 gong1 bei4/的开放二音节单元涵盖了/shi4 ban4/、/shi4 bei4/、/gong1 bei1/以及/bei4 gong4/等多种与/shi4 bei4 gong1 ban4/重叠的激活可能性。因此，考察汉语口语识别中的语音编码的研究或将为该模型提供更为扎实的证据。

5 总结与展望

本综述聚焦口语识别中语音位置编码的新进展，重点梳理了在音位、音节以及句子水平上对

语音位置编码进行探讨的研究。总体来看，以往研究一致认为口语识别采用线性序列编码的加工方式，词汇的语音表征是基于位置固定的序列。但近五年的研究发现拼音语言中口语词汇识别中的语音编码似乎更为灵活，可以脱离其表征中的具体位置单独编码，以往针对语音识别基于“槽”的位置特定编码(slot-based theory)模型受到了挑战。研究者们采用启动范式、视觉情境范式等，通过转置同音节中的两个或三个音位、转置不同音节中的音位、转置句子中的词汇等方式，发现转置前后的词汇、句子可以互相激活，出现了“音位转置效应”、“词汇转置效应”等等。词汇频率、音位距离、词汇距离等自上而下的信息也会影响这种口语识别中的编码过程。总之，虽然声音信号是随时间线性输入的，但我们对语音信号的编码会更加灵活，可以独立于这种固定的位置单独进行。揭示和完善语音信号的编码方式对口语加工模型的建立和对比起着至关重要的作用。

虽然研究者们对拼音文字中的语音编码现象展开了研究，但也存在一些实验材料以及方法上的问题。汉语作为一种表意文字，具有很多特殊性，例如，形音分离，以音节为单位的语音加工单元等等，都可以为我们考察口语编码机制提供更好的素材，也可以使目前的研究结果得到更好的推广。从文献来看，目前这个领域还有很多亟待解决的问题。将来的研究应该围绕揭示听觉语音编码与位置编码的内在机制这个重点，从以下四个方面展开：在本土化上探究汉字词语音位置编码的特殊性，考察口语识别中语音位置编码的神经机制，利用已有和潜在的研究成果来指导不同人群的语言获得与学习过程，构建、改善计算模型并促进人工智能的发展。解决这些问题将会极大促进我们对人类语言认知过程的全面理解。

第一，通过考察汉字词语音编码的特殊性，进一步揭示听觉语音编码与位置编码的内在机制。尽管目前已有研究开始关注口语识别中的语音编码机制，但仍然存在一些尚未解决的问题。例如：(1)拼音语言中对语音位置编码的探讨如何规避形音联结的问题；(2)音节水平的语音位置是否可以采取灵活编码的方式。加之汉字词的众多特殊性也使其加工与拼音文字系统有重要的区别。因此，口语识别中汉字词的语音编码与位置编码的内在认知机理将会是未来心理语言学非常重要的研究课题。

第二，利用脑电技术、功能性核磁共振等技术探索与语音时序位置处理相关的神经机制。尽管使用行为实验以及眼动追踪实验研究口语识别中语音编码的文献丰富，但有关神经机制的探讨相对较少。研究者主要发现了颞上回与语音编码的重要关联(Scott, 2019; Yi et al., 2019)。Yee 等人(2008)报告了布洛卡区受损以及威尔尼克区受损的患者，他们发现额下回的病变使得词首起始音的竞争效应变得非常小，而颞上回的病变则导致该竞争效应比对照组更大。也有研究者发现左侧缘上回和左侧额下回参与了词汇的语音竞争(Prabhakaran et al., 2006; Righi et al., 2010)。探索语音位置序列编码的神经机制将使我们更深一层理解其内在机制与加工过程。因此，有必要通过使用神经科学技术来探讨语音位置编码机制，从生理方面为语音编码过程提供更扎实的证据。

第三，如何利用现有研究结果来指导不同人群的语言获得与学习过程。不管是儿童，还是二语学习者，口语加工都是非常重要的获得途径和方式。但我们在语言学习的过程中都会存在各种各样的问题。例如，在第二语言产生过程中，句子中词序的颠倒是晚双语者常见的错误。因此，如何使用现有与潜在的研究结果与理论开发更合理的方式去训练与干预语言学习者的语言习得过程，以提高其语言加工效率，显得尤为重要。

第四，人工智能已经快速渗透进现代生活的方方面面，在信息技术不断革新的过程中飞速发展，诸如 GPT 等自然语言处理的模型已崭露头脚。在语音处理方面，手机语音助手、车载语音助手以及智能家居等等也趋近成熟。以往关于语音处理的模型也多基于线性序列匹配进行，探索更进一步更高维度的计算模型可以帮助人工智能技术实现更多更快更全面的功能。然而，现有的拼音文字口语加工模型还需改进和完善，关于汉语口语加工的理论模型还处于空白阶段。因此，探讨语音编码机制可以使我们更深入地了解人类处理言语信号的机制，从而为人工智能助手等进一步的发展提供科学依据。

围绕口语识别过程中的语音位置编码机制，本综述旨在澄清语音编码是基于固定位置的序列位置编码，还是独立于语音位置的灵活位置编码。尽管序列位置编码方式和相关理论模型已相

对成熟, 并得到大量研究支持, 但近年来提出的语音灵活位置编码方式仍然缺乏充分的定量支持。尽管已发现了独立于位置的语音编码方式, 但一些研究者仍持有诸如元音位置的重要性等观点。研究者通过启动范式和视觉情境范式等手段观察音位转置和词汇转置效应, 然而, 在实验设计和分析上是否存在系统性方法仍需深入研究。此外, 对于不同语言特征之间的交互影响, 目前尚未详尽探讨。综合而言, 虽然本文在揭示口语识别中语音位置编码方面提供了重要见解, 但研究方法的深化、跨语言比较的拓展以及对影响因素的深入分析等方面仍有待进一步的探索和完善。这些方面的加强还取决于更多研究者对语音编码问题的深入研究, 这将有助于更全面、准确地理解语音位置编码的本质及其在口语识别中的作用。

参考文献

韩海宾, 许萍萍, 屈青青, 程茜, 李兴珊. (2019). 语言加工过程中的视听跨通道整合. *心理科学进展, 27*(3), 475−489.

黄伯荣, 廖序东. (2011). *现代汉语* (上册, 增订五版). 北京: 高等教育出版社.

彭聃龄, 丁国盛, 王春茂, Taft, 朱晓平. (1999). 汉语逆序词的加工——词素在词加工中的作用. *心理学报*, *1*, 36−46.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(38), 419−439.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*(2), 208−216.

Chambers, S. M. (1979). Letter and order information in lexical access. *Journal of Verbal Learning and Verbal Behavior, 18*(2), 225−241.

Chen, J. Y., Chen, T. M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language, 46*, 751−781.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*, 417−430.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, *32*(2), 193−210.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256−266.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, *107*(1), 84−107.

Dahan, D., & Magnuson, J. S. (2006). Spoken Word Recognition. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 249–283). Academic Press.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*, 713−758.

Dufour, S., & Frauenfelder, U. H. (2010). Phonological neighbourhood effects in French spoken-word recognition. *Quarterly Journal of Experimental Psychology*, *63*(2), 226−238.

Dufour, S., & Grainger, J. (2019). Phoneme‐order encoding during spoken word recognition: A priming investigation. *Cognitive Science*, *43*(10), e12785.

Dufour, S., & Grainger, J. (2020). The influence of word frequency on the transposed-phoneme priming effect. *Attention, Perception, & Psychophysics, 82*(6), 2785−2792.

Dufour, S., & Grainger, J. (2022). When you hear /baksɛt/ do you think /baskɛt/? Evidence for transposed-phoneme effect with multisyllabic words. Journal of Experimental Psychology: Learning, Memory, and Cognition, 48(1), 98−107.

Dufour, S., Mirault, J., & Grainger, J. (2021). Do you want /ʃoloka/ on a /bistɔk/? On the scope of transposed-phoneme effects with non-adjacent phonemes. *Psychonomic Bulletin & Review*, *28*(5), 1668−1678.

Dufour, S., Mirault, J., & Grainger, J. (2022). Transposed-word effects in speeded grammatical decisions to sequences of spoken words. *Scientific Reports, 12*(1), 22035.

Dufour, S., Mirault, J., & Grainger, J. (2023). When facilitation becomes inhibition: Effects of modality and lexicality on transposed-phoneme priming. *Language, Cognition and Neuroscience*, *38*(2), 147−156.

Dufour, S., & Peereman, R. (2003). Inhibitory priming effects in auditory word recognition: When the target's competitors conflict with the prime word. *Cognition*, *88*(3), B33−B44.

Frankish, C., & Turner, E. (2007). SIHGT and SUNOD: The role of orthography and phonology in the perception of transposed letter anagrams. *Journal of Memory and Language*, *56*(2), 189−211.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613−656.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E.,

& Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079−1088.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review*, *115*(3), 577−600.

Grainger, J., & Van Heuven, W. J. B. (2004). Modeling letter position coding in printed word perception. In P. Bonin (Ed.), *Mental lexicon: "Some words to talk about words"* (pp. 1−23). Nova Science Publishers.

Grainger, J., & Whitney, C. (2004). Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Sciences, 8*, 58−59.

Gregg, J., Inhoff, A. W., & Connine, C. M. (2019). Re-reconsidering the role of temporal order in spoken word recognition. *Quarterly Journal of Experimental Psychology*, *72*(11), 2574−2583.

Guerrara, C., & Forster, K. (2008). Masked form priming with extreme transposition. *Language & Cognitive Processes, 23*, 117−142.

Gwilliams, L., King, J. R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), 6606.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001* (pp. 159−166). Stroudsburg, PA: Association for Computational Linguistics.

Han, H., & Li, X. (2020). Degree of conceptual overlap affects eye movements in visual world paradigm. *Language, Cognition and Neuroscience*, *35*(10), 1456−1464.

Hannagan, T., Dupoux, E., & Christophe, A. (2011). Holographic string encoding. *Cognitive Science, 35*, 79−118.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, *4*, 563.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in aachine learning. *The Annals of Statistics*, *36*(3), 1171−1220.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), 23−32.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science*, *20*(2), 137−194.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, *38*(3), 245−294.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition 106*(3), 1126−1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *USA*, *106*, 21086−21090.

Liu, Z., Li, Y., Cutter, M. G., Paterson, K. B., & Wang, J. (2022). A transposed-word effect across space and time: Evidence from Chinese. *Cognition*, *218*, 104922.

Liu, Z., Li, Y., Paterson, K. B., & Wang, J. (2020). A transposed-word effect in Chinese reading. *Attention, Perception, & Psychophysics*, *82*(8), 3788−3794.

Liu, Z., Li, Y., & Wang, J. (2021). Context but not reading speed modulates transposed-word effects in Chinese reading. *Acta Psychologica*, *215*, 103272.

Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*, 615−625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1−36.

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148−172). Cambridge, MA: MIT Press.

Marslen-Wilson, W. D., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*(6), 1376−1392.

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101*(4), 653−675.

Marslen-Wilson, W. D., & Tyler, L. K. (1987). Against modularity. In J. L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (pp. 37−62). Cambridge: The MIT Press.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*(1), 29−63.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576−585.

Marslen-Wilson, W. D. (1993). Issues of process and representation in lexical access. In G. T. M. Altmann & R. Shillcock (Eds.), Cognitive models of speech processing: The second Sperlonga meeting (pp. 187−210). Lawrence Erlbaum Associates Publishers.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522−523.

Marslen-Wilson, W. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4*, 55−73.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1−86.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*, B33−B42

Mirault, J., Snell, J., & Grainger, J. (2018). You that read wrong again! A transposed-word effect in grammaticality judgments. *Psychological Science*, *29*(12), 1922−1929.

Norris, D. (1994). SHORTLIST: A connectionist model of continuous speech recognition. *Cognition, 52*, 189−234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*, 357−395.

O'Connor, R. E., & Forster, K. I. (1981). Criterion bias and search sequence bias in word recognition. *Memory & Cognition*, *9*, 78−92.

O'Seaghdha, P. G., Chen, J. Y., & Chen, T. M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, *115*(2), 282−302.

Perea, M., & Lupker, S. J. (2003). Does jugde activate COURT? Transposed-letter similarity effects in masked associative priming. *Memory & Cognition, 31*, 829−841.

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory and Language, 51*, 231−246.

Prabhakaran, R., Blumstein, S. E., Myers, E. B., Hutchison, E., & Britton, B. (2006). An event-related fMRI investigation of phonological−lexical competition. *Neuropsychologia, 44*, 2209−2221.

Qu, Q. Q., Damian, M. F., & Kazanina, N. (2012). Sound-size segments are significant for Mandarin speakers. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 109*, 14265−14270.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive psychology, 7*(1), 65−81.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*(1), 125−157.

Righi, G., Blumstein, S. E., Mertus, J., & Worden, M. S. (2010). Neural systems underlying lexical competition: An eye tracking and fMRI study. *Journal of Cognitive Neuroscience*, *22*(2), 213−224.

Scott, S. K. (2019). From speech and talkers to the social world: The neural processing of human spoken language. *Science*, *366*(6461), 58−62.

Sereno, S. C., Brewer, C. C., & O'Donnell, P. J. (2003). Context effects in word recognition: Evidence for early interactive processing. *Psychological Science*, *14*(4), 328−333.

Toscano, J. C., Anderson, N. D., & McMurray, B. (2013). Reconsidering the role of temporal order in spoken word recognition. *Psychonomic Bulletin & Review*, *20*(5), 981−987.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 394−417.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review*, *8*, 221−243.

Yee, E., Blumstein, S., & Sedivy, J. C. (2008). Lexical−semantic activation in Broca's and Wernicke's aphasia: Evidence from eye movements. *Journal of Cognitive Neuroscience, 20*(4), 592−612.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096-1110.

You, H., & Magnuson, J. S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*, *50*, 871−889.

You, W., Zhang, Q., & Verdonschot, R. G. (2012). Masked syllable priming effects in word and picture naming in Chinese. *PloS one, 7*(10), e46595.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*(1), 25−64.

The mechanism of phonetic position encoding in spoken word recognition

HAN Haibin[1], LI Xingshan[2,3]

([1] *College of Education, Hebei Normal University, Shijiazhuang* 050024, *China*)

([2] *Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

([3] *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: Across various languages, there exists a set of words that retain their meaning even when their phonetic components are transposed. A typical illustration can be found in Chinese with words like "牛黄/niu2 huang2/" and "黄牛/huang2 niu2/". Elaborating on the encoding of such transposable words in the process of language comprehension has become a critically important research topic. Within the field of reading, scholars have been engaged in discussions regarding the mechanisms for encoding word positions.

However, there remains controversy regarding the cognitive mechanisms for phonetic position encoding in spoken word recognition. Early theories posited that phonetic position encoding primarily followed a sequential approach, while recent studies have discovered that phonetic position encoding can be more flexible at levels of phonemes, syllables, and sentences. Future research should delve into questions related to phonetic encoding in spoken word recognition, including cognitive and neural mechanisms, language acquisition, and artificial intelligence. Given the unique characteristics of Chinese characters, such as their spelling-sound dissociation and processing units, subsequent research should pay special attention to the phonetic position encoding in Chinese spoken word recognition.

Keywords: spoken word recognition, phonetic position encoding, Chinese character